DOCUMENT RESUME

ED 155 616 CS 004 131

AUTHOR Carroll, Karen
TITLE A Taxonomic Examination of Seven Reading Comprehension Tests for the Intermediate Grades.
PUB DATE May 78
NOTE 145p.; M.Ed. Thesis, Rutgers, The State University of New Jersey

EDRS PRICE MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Evaluation Methods; Intermediate Grades; *Reading Comprehension; Reading Research; *Reading Tests; *Taxonomy; *Test Construction; Test Reviews

ABSTRACT

The purpose of this study was to evaluate seven reading comprehension tests according to the construct of comprehension which they claimed to measure. The 367 items from these tests were classified by two judges according to the following four levels indicated on a reading comprehension taxonomy: recognition and recall, inference, evaluation, and appreciation. The percentage of items in each category was compared within and across tests. Inference items and recognition tasks together accounted for 81% to 100% of the total items, while evaluation and appreciation items appeared less frequently and were far less developed. The relative weight among items varied from test to test. The following conclusions emerged from the study: most tests contained few questions that related to higher level cognitive processes, judges encountered difficulties in categorization due to differences in definitions, and the low rate of inter-judge agreement in item categorization reflects a need to control the subjective aspects of further analysis. Nine pages of references and ten appendixes are included. (MAI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED155616

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Karen Carroll

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A TAXONOMIC EXAMINATION OF SEVEN READING COMPREHENSION TESTS FOR THE INTERMEDIATE GRADES

A THESIS

SUBMITTED TO THE FACULTY

OF THE GRADUATE SCHOOL OF EDUCATION

OF

RUTGERS

THE STATE UNIVERSITY OF NEW JERSEY

BY

KAREN CARROLL

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

MASTER OF EDUCATION

NEW BRUNSWICK, NEW JERSEY — MAY, 1978

APPROVED: Martin Kling
Martin Kling

John J. Geyer
John J. Geyer

Henry Voos
Henry Voos

ACTING DEAN:
Mary Bach Kievit

CS004131

This thesis is dedicated to my parents,
who first taught me to love learning,
and to my sister,
who is always ready with
encouragement and enthusiasm.

ACKNOWLEDGEMENTS

Appreciation is expressed to my thesis committee chairperson, Dr. Martin Kling, for his interest in this research, and to Dr. John J. Geyer and Dr. Henry Voos for their help in refining the paper. Gratitude is also expressed to Dr. Joseph Zelnick who initially suggested exploring this topic.

Special thanks are extended to Michael Buckman for his interest and insights as the second judge for the study.

Grateful recognition is given to Emily Fabiano of the Oscar K. Buros Learning Resource Center for her encouragement during the final stages of the writing, Gloria L. Johnson of Rutgers Reading Center for her help throughout my program, and to Camille Fernicola for her professional preparation of the manuscript.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF TABLES v

Chapter

I. THE PROBLEM 1

Statement of the Problem 3
Importance of the Study 4
Definition of Terms 4
Limitations of This Study 5

II. REVIEW OF THE LITERATURE 7

Comprehension Theory 13
Testing of Comprehension 33
Subjective and Objective Test Evaluation . . 42

III. PROCEDURES 56

General Procedure 57
Selection of Tests 58
Description of Tests 58
Selection of Taxonomy 65
Description of Taxonomy 65
Application of Taxonomy 68
Treatment of Data 69

IV. FINDINGS AND DISCUSSION 70

V. SUMMARY AND CONCLUSIONS 84

Summary 84
Conclusions 85
Considerations in Taxonomic Analysis 89

REFERENCES 91

APPENDICES

A. Condensed Version of the Taxonomy of Educational Objectives 100

B. Sanders' Preliminary Definitions of Bloom's Categories of Thinking 103

C. Reading Comprehension Question-Response Inventory 104

D. The Barrett Taxonomy--Cognitive and Affective Dimensions of Reading Comprehension 106

E. Taxonomy of Reading Comprehension 112

F. Santos' Classification Scheme for Reading Questions 118

G. Classification of Test Items According to Barrett's Taxonomy of Reading Comprehension 124

H. Problem Items: Reasons for the Difficulty of Placement--Judge A 131

I. Summary of Placement of Sample of Items Showing Item Numbers--Judge B 134

J. Problem Items: Reasons for Difficulty of Placement--Judge B 135

LIST OF TABLES

1. Comparison of Barrett's (1972) Taxonomy of Reading Comprehension with Four Other Taxonomies of Thinking Skills 28

2. General Description of Seven Standardized Reading Tests 59

3. Technical Description of Seven Standardized Reading Tests 60

4. Format of Tests and Items for Seven Standardized Reading Tests 62

5. Publishers' Statements of Skills Measured on Seven Standardized Reading Tests 63

6. Distribution of Items by Category for Seven Standardized Reading Tests 71

7. Percentage of Recognition and Inference Items on Seven Standardized Reading Tests 74

8. Categorization of Sample of Items from Seven Standardized Reading Tests by Two Judges . . . 80

9. Types of Disagreements in Categorization of Sample of Items from Seven Standardized Reading Tests by Two Judges 82

CHAPTER I

THE PROBLEM

Testing has been both a controversial topic and a widespread practice in the field of reading. Buros' Reading Tests and Reviews II (1975) lists 248 reading tests in print as of early 1974. This includes tests in eight categories: general, miscellaneous, oral, diagnostic, readiness, special fields, speed, and study skills.

This proliferation and specialization of tests occurs at the same time that organizations such as the National Education Association recommend caution. In a statement in the October, 1977 NEA Reporter, Ryor urged "a moratorium on standardized tests in general because the tests are misleading at best and provide misinformation to students, parents, teachers and the general public" (p. 3).

Researchers also call for more exact, methodical study of testing in reading because,

> first, current approaches to evaluating reading are seriously deficient and must be improved. . . . Second, current procedures--informal as well as formal--for evaluating reading display basic misconceptions about language. (Wanat, 1977, p. ix)

Questions about what constitutes "reading," what areas of knowledge and skill to test, and the best methods

of testing are reflected in the wide variety of stated purposes, components, and formats that exist in the available instruments. Until these questions are answered, the tests will continue to provide confusing information. Furthermore, if the tests themselves are based on erroneous or incomplete theory, any research that uses scores from those tests as evidence for its conclusions is subject to misleading interpretation. The only possible result from such research is more ineffective teaching and further nonproductive research.

Accurate information about students and programs is needed to design learning situations. As Eller and Farr (1977) pointed out, "A test should provide specific information for making decisions" (p. 29). It must be clear what information the test does yield, i.e., exactly what is measured.

The information gained from a test may be blurred, of course, by many factors: (a) a poor test instrument, (b) inappropriate procedures for administering the test, (c) inappropriate interpretation of the test results, or (d) inappropriate educational application of the test results (Wanat, 1977, p. ix).

This paper will concentrate on the tests themselves rather than their administration or scoring. It will focus on the relationship of the test items to the underlying concept of comprehension. Teachers and

researchers have called for testwriters to give a clear, defensible statement of the comprehension skills they are testing (Davis, 1972; Simons, 1971; Wanat, 1977). The Davis Reading Test (1962) and several standardized batteries (e.g., Iowa Tests of Basic Skills, 1971; Stanford Achievement Tests, 1973) now provide a skills breakdown as part of their item analysis. Research must validate these skill labels for the items. In order to do that, a full, consistent theory of comprehension is necessary; but such a theory is not yet available (Simons, 1971).

The crucial measure of the relationship of the test items to the underlying concept of comprehension is the construct validity of the instrument. Construct validity requires a well-defined theory and a thorough rationale for the individual items. The taxonomy structure has been suggested as a tool in the search for construct validity (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) and will be explored here.

Statement of the Problem

This paper will examine the relationship of Barrett's Taxonomy of Reading Comprehension (1972) to seven major tests of reading comprehension (listed below) by categorizing the items on those tests according to this taxonomy.

The analysis of the classification results will attempt to answer three questions:

1. What differences will the classification results reveal in the distribution of question types among the seven tests?

2. How frequently will a judge be unable to place an item in one of the available taxonomy categories because the category is too general, too specific, or inappropriate in the opinion of the judge?

3. How frequently will two judges make identical placements of items in the taxonomy categories?

Importance of the Study

Since comprehension is often defined in terms of test results, a thorough understanding of those instruments is essential. This study will build on previous efforts to evaluate construct validity and to analyze tests according to the types of questions they contain. Such data may be useful in the effort to design more useful tests, to define reading/thinking skills, and to build a defensible theory of reading comprehension.

Definition of Terms

The central term for this paper--comprehension--will be discussed fully in Chapter II.

Comprehension test: in this paper, one of the seven standardized tests or subtests of comprehension

originally contained in the Anchor Test Study (Bianchini & Loret, 1974). These tests include:

California Achievement Tests, 1970

Comprehensive Tests of Basic Skills, 1968

(1968-73 edition used in this study)

Iowa Tests of Basic Skills, 1971

Metropolitan Achievement Test, 1970

SRA Achievement Series, 1971

Sequential Tests of Educational Progress, 1969

Stanford Achievement Tests, 1964

(1973 edition used in this study)

Construct validity: the degree to which a test can be related to the abstract psychological variable it is designed to measure (Stanley & Hopkins, 1972, p. 111-112).

Taxonomy: a classification system which orders phenomena in ways which reveal some of their essential properties as well as the interrelationships among them. The order of the terms in a taxonomy must correspond to some "real order among the phenomena represented by the terms" (Bloom et al., 1956, p. 17).

Limitations of This Study

This study has four major limitations:

1. The taxonomy--the placement of each item will tend to reflect the degree of specificity which exists

in the taxonomy. Categories may be too general, too specific, or otherwise inappropriate for certain items.

This limitation was dealt with, in part, by allowing the judges to place an item outside of all the given categories if that placement could be justified. The judges were also asked to provide a brief statement of their conflicting reasoning when placement of an item was difficult.

2. The comprehension tests--decisions made by the testwriters and publishers will predetermine the number and complexity of the comprehension skills required by the tests. It is a general consensus of the field that many of the tests have been constructed without reference to a theory which would provide a focus for each question and a balance of questions within the test (Simons, 1971, p. 317).

3. Rater bias--the placement of items within the taxonomic categories is a subjective evaluation. The background of the judges in the area of comprehension, their interpretation of the categories, and their personal bias about individual items will influence each decision.

4. The number of judges--the limited number of judges in this study prevents the identification of trends in the categorization of the items on these seven tests.

CHAPTER II

REVIEW OF THE LITERATURE

This paper will view comprehension as a theoretical construct measured by seven reading comprehension tests. Therefore, a review of the literature on comprehension theory and comprehension testing is necessary.

Theory must be built on unambiguous definition of terms (Hill, 1977-1978). A system of measurement must also begin with "a clear and consistent definition of the things being counted" (Anderson, 1972, p. 145). Comprehension is a central concept in the theory of reading, but it has steadfastly resisted definition.

Auerbach (1971) has traced the historical sequence of the study of comprehension. She found three approaches which developed successively but which are each part of the current effort to clarify the concept:

1. Philosophical investigation was based on intuition and logical analysis of the reading process and extended from the work of Locke in 1697 to that of Wittgenstein in 1958.

2. Experimental psychology focused on the testing of hypotheses and the specifying of goals and criteria for reading performance. The observed behaviors of large

groups of readers formed the empirical phenomena studied by figures such as Huey (1908/1968), Thorndike (1917), James (1928), and Skinner (1937).

3. Statistics was an approach developed for better objective evaluation of data either in the study of the comprehension process itself or in the correlation of other factors with comprehension scores. Singer (1964), for example, extended the substrata-factor theory developed by Holmes (1948, 1965) into a statistical model of reading. Davis' list of subskills in reading was statistically derived from data on student responses to carefully prepared question sets (1941).

Each of these three approaches has contributed to knowledge about the basic process, but a definitive explanation of the full process is not yet available. In 1969, Farr observed "there is still a lack of understanding about the basic aspects of reading comprehension" (p. 56). Simons (1971) felt that this was due to more than a lack of information: "The problem is not that more facts are needed but that it is impossible at present to give a coherent explanation of the facts that already exist" (p. 355).

Cleland (1965) credited Grey in 1917 with being the first to use the term comprehension to denote the obtaining of meaning through reading. In that same year, Thorndike (1917) proposed his classic definition:

> [Comprehension is] a very complex procedure, involving a weighing of each of many elements in a sentence, their organization in the proper relations to one another, the selection of certain of their connotations and the rejection of others, and the cooperation of many forces to produce the final response. (p. 329)

Subsequent work referred to the steps of Thorndike's complex procedure as separate skills and began a long effort to explain them and their interrelationships. The term "comprehension skills" is now one of the most ubiquitous and troublesome in the field.

Simons (1971) argued that much of the current confusion arose from three different interpretations of the term. "Skills" has been used to refer to (a) uses of comprehension (i.e., behavior), (b) procedures for teaching comprehension, and (c) psychological processes involved in reading comprehension.

Drawing this distinction gave a clearer picture of the relationship between the inner workings of the mind and the overt reading performance. Harker (1975) saw the process and the behavior as clearly different: "The performance of skills is merely the manifestation of this extremely complex inner process . . . involving the cognitive processing of language" (p. 6).

The most difficult issues in testing arise from this distinction. Spache (1963), for instance, noted the inexact labeling in tests of comprehension: "The distinction between observing the process of comprehension

and noting the end-results of the process is significant" (p. 84). He stated that intellectual processes (i.e., verbal ability, reasoning, and general intelligence) appeared to influence comprehension but that neither the processes nor their exact relationship to comprehension were fully defined.

Bloom et al. (1956) have drawn a similar dichotomy between intellectual skills and intellectual abilities. In order to demonstrate a skill, the reader needs only that information available from a fund of common knowledge. The real emphasis is on the method of dealing with the problem, i.e., organizing and reorganizing information. On the other hand, to demonstrate an ability, specific knowledge must be called up to handle a new problem. In this case, the task demands both the memory of and the choice of the correct information to apply to the situation.

The construct which Bloom et al. labeled ability seems most closely allied with Simon's psychological processes while their skill has some characteristics of Simon's uses of comprehension.

However, after drawing the distinction between a skill and an ability, Bloom did not use it in the taxonomy. They pointed out that it is the prior experience of the student which determines whether a problem is new and the information is available and that prior experience

of each individual cannot be gauged by the test writer. The omission is regrettable, and future research must find some way to account for prior experience in measuring comprehension.

Another serious issue is whether to define comprehension as a passive or an active process. Either the reader receives "the meaning" from a passage (Carver, 1977-1978; Gough, 1972) or creates "a meaning" that depends on personal past experience (Goodman, 1976). Rystrom (1977) drew each of these positions out to their logical conclusion and found both unable to account for some of the phenomena observed in reading.

He suggested instead that reading is a matrixing event between the reader and the print. A matrix is a set of items simultaneously influencing and being influenced by each other. In Rystrom's explanation, the comprehension which results from reading is a matrix which incorporates elements of both the author's and the reader's own matrix of knowledge into a structure that is, in turn, open to further modification.

This is similar to Wittrock's (1977) conception of a process that uses the print to stir memories of distinctive personal experience and generates meaning by processing those memories semantically.

In both cases, the prior experience makes the process different for each person, but communication is

possible in spite of that. A robust theory of comprehension is needed to abstract similarities in each person's tactics and to explain the individual differences in terms of a definition of the basic process.

Two efforts to summarize current knowledge on reading have given priority to defining comprehension: the USOE Literature Search Project titled The Literature of Research in Reading with Emphasis on Models (Davis, 1971) and the 1975 NIE Conference on Studies in Reading. The first pointed to "an interest in developing models of reading and more nearly adequate definition of it" as a major theme in recent reading research (p. 2-2). The NIE panel set the following goal for its conference: "To propose approaches to establishing a knowledge base on which more successful teaching of reading comprehension skills might rest" (Miller, 1975, p. 729).

While aiding in the design of better instruction, research on comprehension can also contribute to the understanding of the process of cognition. Reading and thinking are interdependent processes (Betts, 1961; Gibson & Levin, 1975; Kerfoot, 1965; Stauffer, 1969), and development in the two fields will be reciprocal.

However, Weaver (1977a) pointed out that basic tools for this exploration are limited:

> The difficulty for examining the reading process is that the accurate instrumentation, i.e., the measurements, . . . is not available to the reading

> researcher. Theory, however in proposing that which one needs to know to structure a body of data, encourages the development of proper measurement. (p. 59)

The remainder of this literature review will focus on comprehension theory and testing.

Comprehension Theory

The "new perspective" on reading comprehension which Simons requested in 1971 included renewed attention to its theoretical base. Theory is important because it organizes known facts about a phenomenon revealing any gaps in the data or the logic of the presentation. These gaps indicate the areas in which further research is most essential.

Kerlinger's (1965) definition of theory stressed its function in clarifying the available knowledge:

> A theory is a set of interrelated constructs (concepts), definitions, and propositions that present a systematic view of phenomena by specifying relations among variables with the purpose of explaining and predicting the phenomena. (p. 11)

Although some empirical psychologists argue that a theory may bias their approach to data collection (Skinner, 1950), the consensus in the literature is that theoretical direction could speed progress toward an explanation of the reading process (Hill, 1977-1978; McCollough, 1976; Simons, 1971). Weaver (1977a) said it best:

> The research is expedient, non-theoretical, and

> therefore, non-accumulative in its impact, i.e., it is not mapped into a body of tested principles. . . . Theory must tell us what evidence we need. (p. 53)

Weaver (1977a) called for a direct relationship between theory and research. He felt the task of a reading theory is to relate aspects of reading to one another. The task of research is to investigate phenomena empirically, to seek actual measures of the relations involved, and to correct and illuminate theory (p. 62).

Why has the massive body of reading research failed to 'illuminate the theory?' Wardhaugh (1969) believed that researchers were being distracted from the direct study of reading into study of its correlates: IQ, attitude, oral reading skill, etc. (p. 4).

Smith (1973) suggested that the confusion persisted because the reading field lacked the psycholinguistic "distinction between the reading process and reading instruction, between theories of what goes on in the head and beliefs about what should go on in the classroom" (p. 5).

However, Kerfoot (1965) stated that efforts to describe what 'goes on in the head' during reading had serious flaws: "Two sources which have contributed to the confusion are: the relationship of comprehension to intelligence and the use of terminology as personal convenience (p. 253).

Hill (1977-1978) suggested that theory could deal with these problems by using two devices: (a) definitions, and (b) constructs. The definitions must be rigorously examined for clarity and directness: "The definitions should be subjected to reductional processing successively re-defining its terms in order to reach a point where no ambiguity remains about the referents of the original term" (p. 72).

The constructs would define the exact relationship of various aspects of comprehension with each other as well as the relationship of the construct of reading comprehension with other constructs (e.g., intelligence). The relationship would be open to proof through testing. "A construct . . . consists of an inquiry-conscious concept or generalization which represents the relationships between variables, events, or processes" (p. 73).

The two types of constructs are empirical and hypothetical. Hill explained that empirical constructs take the form of a "when-then" statement which can be demonstrated and verified. Hypothetical constructs use an "if-then" statement to show a presumed relationship, subject to further testing (p. 78).

In listing criteria for a theory of reading, Simons (1971) included this emphasis on precision. He defined a theory as

> a group of hypotheses . . . which meets these minimum requirements:
>
> (a) A theory must be falsifiable, i.e., it must be clear what kind of data will confirm or deny it.
>
> (b) A theory must be perfectly explicit, i.e., all the terms must have explicit definitions.
>
> (c) A theory must be comprehensive, i.e., it must include a description of the entire system it purports to explain.
>
> (d) A theory must possess descriptive adequacy, i.e., it must describe the facts accurately.
>
> (e) A theory must be internally consistent, i.e., none of its parts may contradict one another. (p. 340)

In such a theory, terms would have to be unambiguous, and contradictory research results would have to be reconciled, or they would force an enlargement of the theory so that it could remain both comprehensive and internally consistent.

Hill found few attempts in the literature to discuss reading theory at this level of sophistication. In addition to Simons, he credited Gibson and Levin (1975), Huey (1908/1968), Jenkinson (1969), Robinson (1968), and Schutz (1967) with contributions in this area. Each of the major viewpoints in reading was represented in this list--from Huey's early, and still viable--discussion of reading as information-processing to Simon's psycholinguistic perspective and Gibson and Levin's theory of perceptual learning and cognitive strategies in reading.

These theories dealt with different units and aspects of the reading process. Synthesis of accurate research will eventually reveal how the data from one

line of research can be consistent with findings in another.

One major hurdle in theory formation is the choice of the level on which the theory will operate. Weaver (1977a) felt that research has suffered because too often the choice was an arbitrary one. He advocated that a theoretician "delimit the domain . . . to cover a significant segment of the process while maintaining a firm connection with the tested and the testable" (p. 52).

Both of these recommendations--a significant domain and an empirical base--were among Hill's requirements for formal theory statement. Formal theory, or Level 5, was the highest step of a hierarchy by which he categorized explanations of reading:

> Level 5: Complete theoretical statement concerning an identified and significant theoretical phenomenon of reading--
> (a) which includes a satisfactory description of the significance and attributes of the phenomenon;
> (b) which clearly specifies the relationship between its essential components or units;
> (c) which utilizes a minimum of primitive terms in developing a defensible constitutive, logical, and empirical definition of those theoretical units;
> (d) which generates a series of field useful and meaningful predictions; and
> (e) from which supporting empirical evidence has been and is being generated. (p. 79)

Classification

Traditionally, two tools have been used to organize the components of reading for inclusion in a theory: classification schemes and taxonomies. Classification

involves merely separating items into categories, while a taxonomy requires the ordering of category members to show precise relations and properties (Bloom et al., 1956).

A classification scheme for comprehension had much to recommend it as a basic tool but only for the early stages of research. The definitions used in classification are not exact enough for formal theory. An example of its use was Spache's classification of the reasoning required by various test questions (1963). Because the intellectual processes were not fully defined, he declined to rank the relative difficulty of the items or specify their relation to one another.

Körner (1975) provided a thorough discussion of classification theory in the Macropedia of the Encyclopaedia Britannica which stated the conditions necessary for a classification. Classification of a domain "coincides with what in the mathematical theory of sets, is called a 'partition,' i.e., [the subsets] are mutually exclusive and jointly exhaustive" (p. 691). In other words, the subsets are divided according to some criterion. They do not overlap; but taken together, they include every member item.

In classification schemes for comprehension processes, the primary problem is to prove that the scheme is exhaustive, i.e., that it accounts for all of the involved

processes. At present, no definitive study has dealt with this; but, instead, the field has relied on logic and intuition in the effort to reach a consensus on a complete list of comprehension skills.

The former requirement, i.e., ensuring that the subsets of comprehension skills are mutually exclusive, is also hard to satisfy. Work in that area has included factor analytic studies (Holmes, 1965; Singer, 1964) and efforts to construct "clean" tests (Calfee, 1976).

These problems in classification suggest the added difficulties involved in constructing taxonomies which show clearly the relationships among the skills.

Two different approaches to classification are seen in science and social studies; traces of both are found in current reading research.

In science, classifications are based on the frequency and probability distributions of variables within populations. Thus, the empirical data, not a set of constant criteria, determine the classes. The field of reading is dealing with a covert process rather than concrete objects available for handling or visual examination. Therefore, researchers often work with a body of statistical data which represents the process (Holmes, 1965). The data forms the "skills" to be studied. Stennett, Smythe, and Hardy (1975), however, found many of the classic statistical methods for representing

subskills to be misleading and unproductive for further research.

In social science, the place of theory is often taken by a classification model which can be modified according to inquiry results. In reading, such verbal or diagrammatic models have been able to visualize the process in steps or components with varying success (Davis, 1976; Harker, 1973; Singer, 1962). In reading, a model is often built to illustrate an early stage of a theory in hope that it will suggest revisions for that theory.

The Britannica discussion cautioned that models do have many characteristics which simply follow from a sense of what is reasonable with a given situation. The reading field must be careful to keep models closely tied to theory because it has suffered already from an overabundance of intuition and a lack of hard data in its studies of comprehension (Simons, 1971). The testability of models is one way to tie theory to empirical proof.

This discussion began by explaining that a classification did not demand specific hierarchical relations among items in each category. The editors of the Macropedia, however, observed that no scientific classification was ever completely free of theoretical assumptions although these did not need to be stated (p. 692).

Bloom et al. (1956) felt this laxity made

classification schemes too inexact to reflect the precision of a theory. They chose the taxonomy format instead.

Taxonomy

The original purpose of the taxonomy developed in 1956 was to facilitate the exchange of information on testing by members of the American Psychological Association. Standard terminology and classification of phenomena were needed for efficient communication.

The taxonomy provided a hierarchical structure for the range of possible educational goals showing the relationship of each objective to the others. The six ascending levels of the taxonomy were: (I) Knowledge, (II) Comprehension, (III) Application, (IV) Analysis, (V) Synthesis, and (VI) Evaluation.

The authors attempted to base their taxonomy on a comprehensive psychological theory. Recognizing that this theory was still in a developmental stage, they called for further research on a larger synthetic theory of learning to support their structure of skills. That theory is still being sought today.

Bloom et al. aimed to reflect in the taxonomy the order of the actual phenomena. However, they also recognized that the taxonomy was limited by the fact that it placed educational priorities ahead of logical

and psychological ones. Therefore, the educational behaviors in the taxonomy reflect curriculum categories. While they were held consistent with psychological knowledge of the time, they may not mirror the exact sequence of the mental processes a student uses to produce those behaviors.

The nature of the mental processes was in dispute in 1956 when the Taxonomy was written; 20 years later, the dispute still has not been resolved.

> At the present time, a heated controversy exists about whether reading should be taught as a holistic process or through a subskill approach. Those who favor the holistic approach (Goodman, 1972; Smith, 1973) claim that by fractionating the reading process into subskills, the essential meaning deriving aspect of reading is destroyed. In addition, the advocates of the holistic approach claim that when intercorrelations among subskills are computed, the correlations are so high as to suggest that reading is but a single process, and consequently, should not be fractionated into subskills.
>
> On the other hand, those who favor the subskill approach point to the accumulated evidence on complex human learning, which strongly suggests that prior to mastering a complex, higher-order skill, numerous subskills must first be acquired (Gagné, 1962; Hilgard & Marquis, 1961, p. 127). (Samuels, Begy, & Chen, 1975, p. 74)

The arrangement of the taxonomy, of course, implied a gradual movement from simple skills to complex. In reality, answering a test item could involve the use of several subskills simultaneously, of different subskills by different students on the same item, or of different subskills by the same student on different administrations of the test. Clymer (1968) observed: "As is true with

all taxonomies, the orderly presentation of the categories may suggest a greater precision than the classification system really possesses" (p. 19).

While Bloom et al. did not claim that the taxonomy represented a lock-step progression of either skill use or skill development, they felt that there was some evidence to support the general organization.

> Our attempt to arrange educational behaviors from simple to complex was based on the idea that a particular simple behavior may become integrated with other equally simple behaviors to form a more complex behavior. (1956, p. 18)

They called for experimental verification of two trends which they saw as evidence for the hierarchical order in the taxonomy:

1. Questions requiring higher-level thought were answered in a correct way less frequently than those requiring lower-level thought.

2. More complex behaviors were accompanied by more awareness on the part of the individual that the thought processes were taking place. Students could describe them; whereas, lower-level thought was automatic and more difficult to articulate.

Roberts (1976) listed six studies which support the hierarchy of skills which Bloom et al. proposed (Ayers, 1966; Hunkins, 1968; Kropp, Stoker, & Bashaw, 1966; Scriven, 1967; Smith, 1970; Stanley & Bolton, 1957).

She also reported one study which failed to validate Level III, Application (Klein, 1972) and one which concluded that Levels IV (Analysis) and V (Synthesis) may not be cumulative results of the lower level behaviors (Madeus, Woods, & Nuttall, 1973). Stedman (1973) was not able to show an increase in difficulty between questions at the Knowledge level (I) and the Comprehension level (II) or between tasks of Application (III) and Analysis (IV). Roberts' own study found a "weak link" between Level I (Knowledge) and Level II (Comprehension) implying that these are different types of thinking. She also determined that questions for the Application level (III) are particularly difficult to devise (p. 18).

However, one evidence that the levels of the taxonomy are related is the fact that these test results "clearly indicate that learning took place at all levels of thinking when materials and teacher efforts were directed at upper levels only" (p. 18).

Stoker and Kropp (1964) also tested the claim of hierarchical structure for the taxonomy. They prepared two tests consisting of equal numbers of items in each of the major categories of the Taxonomy. Identification of an item as belonging to a specific category rested: (a) on the intuition of the item-writer, (b) on discussions with other project staff members, and (c) on

categorization of the items by doctoral students to confirm the classification.

Stoker and Kropp (1964) found disagreements on placement of items which were related to the two issues already discussed: the use of prior knowledge and alternative processing.

> The item writer and the item judge have different notions about the competencies and problem-solving methods used by the students. . . . First, the item writer might prepare an item for an upper level of the Taxonomy and the rater believes the student will know the answer to the item because of prior knowledge. Second, the item writer might prepare an item to evoke behaviors regarded as "Evaluation," for example, but the student might answer the question by systematically eliminating distractors; thus, the intended process differs from the obtained process. (p. 40)

The majority of placements did show inter-judge agreement in spite of this.

After administration of the test, category levels and mean scores for the items generally showed an inverse relationship indicating that items in the higher levels were more difficult to answer. On one test, results from the Evaluation items did not support this conclusion.

A further check of the items for each level using simplex analysis indicated that the Knowledge, Comprehension, Application, and Analysis levels, respectively, were of increasing complexity.

The expected factor structure underlying the Taxonomy, however, was not confirmed. Knowledge and

Comprehension did appear related; and as the combined scores on these two subtests increased, a similar rise was noted on the other four subtests except Evaluation.

In particular, as Kropp, Stoker, and Bashaw (1966) had pointed out, the interaction of stored content and the process of recall was troublesome.

> The majority of the factors which were extracted were mixtures of content and process. The presence of these mixed factors tends to discredit the hypothesis [of hierarchical structure]. . . . Nevertheless, certain patterns did emerge which suggested somewhat systematic changes over process levels and grades. However, no claim was made about them because of the obvious need for more refined analytic techniques and data from more refined taxonomy-type tests. (p. 168)

One limitation noted in this study was the fact that the responses to items may have been produced by processes other than those which the items were intended to evoke (p. 171). The research team suggested analysis of transcripts of oral problem-solving sessions as an alternative, if time-consuming, means of tracing the processes used by subjects.

All of these difficulties in validation and use have also been encountered by taxonomies which followed this pioneer effort. Bloom's Taxonomy had been written to describe educational outcomes in a broad context, i.e., to apply to the many subject areas and activities that can be part of a learning situation.

Barrett (1968, 1972), Guszak (1967), Sanders (1966),

and Santos (1976) each adapted the basic form for a particular purpose by adding, deleting, subdividing, or combining categories.

Sanders wanted to encourage teachers to ask "questions that require students to use ideas, rather than simply to remember them" (p. 2).

Guszak aimed at improving oral questioning techniques in reading circle instruction in elementary classrooms.

Barrett's (1972) goal was "to provide teachers with an understandable and manageable framework for planning teaching and evaluating in the area of reading comprehension" (p. 58). He revised his set of categories between 1968 and 1972, eliminating a Reorganization category which originally occurred between Literal and Inferential Comprehension. The 1972 version was used in this paper. This decision will be explained in the following discussion.

The Santos Classification Scheme for Reading Questions was proposed as a tool for demonstrating "the comprehension behaviors required for success in content areas" (p. 175). It is essentially an inventory of the basic types of questions which are believed to stimulate various levels of thinking.

The structure of the five taxonomies is presented in Table 1. A fuller statement of each appears in

TABLE 1

COMPARISON OF BARRETT'S (1972) TAXONOMY OF READING COMPREHENSION WITH FOUR OTHER TAXONOMIES OF THINKING SKILLS

Barrett (1972)	Guszak (1967)	Santos (1976)	Sanders (1966)	Bloom, et al.[a] (1956)
LITERAL COMPREHENSION recognition recall	RECOGNITION RECALL	RECOGNITION or RECALL	MEMORY recognition recall	KNOWLEDGE
INFERENCE	TRANSLATION	TRANSLATION	TRANSLATION INTERPRETATION	COMPREHENSION translation interpretation extrapolation
	CONJECTURE	APPLICATION	APPLICATION	APPLICATION
		ANALYSIS	ANALYSIS	ANALYSIS
	EXPLANATION	SYNTHESIS	SYNTHESIS	SYNTHESIS
EVALUATION	EVALUATION	EVALUATION	EVALUATION	EVALUATION
APPRECIATION	–	–	–	AFFECTIVE DOMAIN

[a]Bloom, Englehart, Furst, Hill, and Krathwohl.

Appendices A through F. Table 1 is based on one devised by M. Smith (1977) for her discussion of taxonomies and reading comprehension questions. Considerable agreement was found among the systems at the lowest level of questions which basically refers to recognition and recall of material. Some slight difference in terminology is observed for this level: Literal Comprehension (Barrett), Knowledge (Bloom), and Memory (Sanders). Three of the taxonomies (Bloom, Sanders, and Santos) which differentiate the upper levels of thought use identical labels: Application, Analysis, Synthesis, and Evaluation.

Barrett added a final category of Appreciation which required students to articulate an emotional and aesthetic sensitivity to the tools a writer uses in a selection. This is a subjective statement of effect that is different from the use of external or internal standards in evaluation questions. Bloom, of course, relegated these types of emotional responses to the affective domain and its separate taxonomy.

The area of most inconsistency from one taxonomy to another is the second level which moves beyond literal comprehension. These are the categories that are based on inference. Barrett referred to the entire level by that name. This eliminated the need to differentiate translation (mentioned by each of the other four authors) from other types of inference: explanation, conjecture,

interpretation, and extrapolation. The basic Inference called Translation by Santos is simpler than Guszak's Conjecture or Explanation which also involve Application, Analysis, and Synthesis. The problem appears to lie with the process which Barrett dropped when he revised the 1968 Taxonomy: Reorganization.

> Reorganization requires the student to analyze synthesize, and/or organize ideas or information explicitly stated in the selection. To produce the desired thought product, the reader may utilize the statements of the author verbatim or he may paraphrase or translate the author's statements. [Italics mine.] (p. 10)

The freedom to paraphrase or translate makes this process clearly different from Recall; so in the new version, it must be absorbed into the Inference category. However, a wide range separates the translation of a single word from the performance of a complex analysis or synthesis of implied information to answer an item.

Inference is the central thought process which must be clarified. Even in assigning simple meaning to word forms, the learner is drawing a connection, i.e., an inference, that meaning is in some sense equivalent to the word form.

While Sanders and Guszak made practical distinctions between the simple and complex classroom tasks of Inference, Barrett kept the full range of these thought processes together under a single label.

One of the stated purposes of this paper was to contribute to the knowledge base for the formation of a

theory of reading comprehension. Therefore, the more comprehensive definition of inference was preferred; and Barrett's Taxonomy was chosen. It may result in a different picture of what the questions on comprehension tests are actually requiring test takers to do. The traditional complaint has been that tests and teachers ask too many literal questions and ignore the higher-level thought processes (Barrett, 1972; Guszak, 1967; Sanders, 1966). It may be that some "literal" questions are actually inference items which need to be better understood and used by test writers and teachers.

Another problem with categories in the taxonomy is the relationship of more complex behaviors to simpler ones. Bloom stated that the apparent cumulative structure needed verification (p. 19). Barrett warned that the progression was flexible according to the learner and the situation and must allow for some overlap among the categories (p. 58). Santos called the order of the levels "approximate" (p. 175) but presented a clear-cut chart (see Appendix F) demonstrating "the interrelatedness and increasing complexity of comprehension behaviors as questions move from Recognition and Recall to Evaluation" (p. 182).

Two explanations of complex behaviors are possible (Barrett, p. 16):

1. A complex behavior may be simply a sum of the

component behaviors, or

2. A complex behavior may be more than the sum of simpler actions (i.e., similar to a Gestalt view of a whole vs. its component parts).

Barrett recommended that items be categorized according to "primary intent of the task" (p. 58). This intent is subjectively determined and may be disputed. Sanders, however, stressed that "difficulty in classifying any question is no detraction from the quality of the question" (p. 8).

Sanders suggested that a behavior be inserted into the Taxonomy at the point of highest complexity although it might also exist in lower levels (p. 10). Bloom et al. (1956) acknowledged the fact that broad statements of intended behaviors may be especially hard to classify (p. 46). None of the authors specified how the higher levels would acknowledge the "something more" of complex skills from a Gestalt point of view.

Using a taxonomy to analyze reading behavior reaffirms Thorndike's original statement that "reading is a complex procedure" (1917, p. 329). His definition of comprehension grew out of a concern with testing procedures in reading. The final sections of this literature review will deal with the testing of comprehension.

Testing of Comprehension

Comprehension has been defined in various ways; but in practice, it has been most often equated with performance on reading tests (Auerbach, 1971; Goodman, 1976). Auerbach also found that to a large extent the tests reflect their author's conception of comprehension. Thus, it is not surprising to find substantial variation among reading tests in terms of what aspects of comprehension are measured (Pyrczak & Rasmussen, 1974).

The large number of instruments which contradict or duplicate one another is a direct result of the lack of a theory base according to Simons (1971). This is not to say that they are all carelessly constructed.

Mavrogenes, Winkley, Hanson, and Vacca (1974) listed four types of reading tests: (a) the survey test, (b) the analytical test, (c) the diagnostic test, and (d) the special test. The typical instrument consists of passages followed by multiple-choice questions or presents a completion task based on the paragraphs. Often it is part of a battery; and generally, its items are arranged from easy to difficult.

In general, tests have been developed according to a statistically sound plan. As Auerbach outlined it, the first step is to administer experimental test items, as well as an IQ test, to large population samples at many grade levels. Based on this data, test items are

chosen because (a) they discriminate empirically among older and younger pupils, and among those with high and low IQs, and (b) the questions as a total group have empirical difficulty scores which reflect a normal distribution; i.e., 23% are difficult, 54% are of average difficulty (passed by 70-80% of the population), and 23% are easy. (Note: Empirical difficulty has been defined as the proportion of the population which answers an item correctly.) Standardization is then carried out with these selected items, and norms are calculated which allow the ranking of an individual's performance in relation to the standardizing population (Auerbach, 1971, pp. 21-23).

This procedure does not deal with the question of the extent to which the resulting test actually relates to a coherent theory of comprehension. Kingston (1960) pointed out:

> Most of our measures have little or no theoretical basis but rather reflect both eclecticism and empiricism in design and content. Any item which yields a statistical coefficient which is reliable or valid is utilized. (p. 119)

To correct this situation, Goodman (1976) listed unequivocal requirements for test construction:

> Building a test requires much more than administering it to various standardization groups to establish statistical norms as well as statistical reliability and validity. Building a test requires a theoretical model of the competence to be examined, based on adequate research and knowledge. Subtests, items and tasks must relate to reality through this model. (p. 94)

Farr and Anastasiow (1969) have observed that individual subtests in these measures do not possess sufficient validity for the subskill they allegedly isolate. Instead, however, their overall scores do provide a global measure that is reliable and valid for estimating the position of the whole class in comparison to the large standardizing population. The value of the subscores remains in question.

Spache (1963) had also expressed concern about the validity of separate subtests:

> There is very little evidence that current tests are true measures of different reading comprehension performance or that they measure different types of thinking, no matter what labels the subtests bear . . . . They merely offer labels for some of the outcomes of the comprehension process. (p. 62)

The labels assume a reality of their own; and once those results are utilized in further research, the problems are compounded. For example, factor analysis studies of comprehension may predispose their results toward the skill labels on the tests they use to generate data.

Simons (1971) contended that the labels may also mask other factors being measured at the same time, such as memory, prior knowledge, vocabulary level, test-taking skills, etc.:

> For this reason traditional reading comprehension tests lack construct validity (Cronbach & Meehl, 1955); that is, no clear relationship has been developed between the test and the underlying concept: comprehension. (p. 347)

Lack of construct validity is one of the major limitations of standardized tests. Spache (1963) has provided a thorough listing of some other problems with these tests:

1. The inherently artificial nature of interrupting the process in order to test it.

2. The basic complexity of the comprehension process.

3. The illusion that one test measures all aspects of a student's reading ability or a school's reading program.

4. The reduction of a multifaceted ability to a single score.

5. The limits placed on reliability by the brevity of the subtests.

6. The effect of the element of speed on reliability.

7. The false impression of accuracy because of the statistical format of the results.

8. The use of hypothetical norms for performance at different levels.

9. The variation among tests in format and conditions of testing, e.g., access to selections while answering items.

Recognizing these limitations, Spache pointed to some informal measures which could demonstrate growth in

comprehension:

1. Increasing accuracy in answering questions of various types.

2. Accuracy in material of increasing difficulty.

3. Improving rate.

4. Increasing amounts of reading matter handled efficiently at one time.

Some of these factors are reflected in existing reading tests, but evidence of progress for an individual would require a long-range testing program with consistent instruments.

Wanat (1977) included more specific criticisms of test administration procedures in his list of major weaknesses found in four widely used reading tests:

1. Tests fail to include important skills.
2. Tests assess skills irrelevant to reading by including tasks not related to reading.
3. Test instructions telling the child what to do are not clear.
4. The tests may confuse intelligence with reading. . . .
5. The test may confuse correlation with causation. . . .
6. The test may have an irrational scoring procedure. . . .
7. The test may be an inefficient way of gathering information since some tests require considerable time. . . .
8. Items in subtests measure such diverse skills that it is hard to ascertain what is actually being assessed.
9. Subtests have too few items to give reliable information.
10. Subtest scores may give information that has no instructional value. . . .
11. The test may provide inadequate instructional

> guidance by failing to relate test outcomes to specific instructional procedures. . . .
> 12. Subtests making up the tests do not measure what they claim to measure. . . . (pp. vii-viii)

The work by Davis (1941, 1972) on subskills in reading was a continuing effort to work with tests at a level of educational behaviors similar to Bloom's. Of the nine operational skills which Davis proposed, five fit his criteria for stability and order. The five emerged as components of comprehension on a factor analysis:

1. Knowledge of word meanings
2. Verbal reasoning
3. Sensitivity to implications
4. Ability to follow the structure of a passage
5. Ability to recognize literary techniques

However, when Thurstone (1946) reevaluated the statistical procedures which Davis had used, he found support for only one general factor. Hunt (1957) isolated two skills from the same data: word knowledge and paragraph comprehension; and Alshan (1964) could not disprove the hypothesis that the five skills were interrelated. In his 1972 work, however, Davis did not substantially change his original conclusions in spite of the statistical problems involved in interpreting the data.

Berg (1973) summarized the work in this highly technical area:

> A rough estimate of the number of factors that researchers suggest can be measured independently . . .

> [is] a number somewhere between two and five. . . . Yet a review of reading tests turns up 70 to 80 factors that various tests implicitly claim to measure. (p. 30)

The two to five independent skills are generally of a global nature such as paragraph comprehension or verbal ability. Research is needed to resolve this discrepancy between testing practice and the fragile evidence which supports comprehension theory.

Factor analysis was only one of the procedures recommended by Anastasi (1966) to confirm construct validity. She also suggested checking for a developmental match with IQ scores and the manipulation of various experimental variables to clarify the nature of reading ability. She referred to Campbell and Fiske's (1959) point that

> in order to demonstrate construct validity we need to show not only that a test correlates highly with other variables with which it should correlate, but also that it does not correlate with variables from which it should differ. The former is described as convergent validation, the latter as discriminant validation. (p. 308)

In its official Standards for Educational and Psychological Tests (in Buros, 1974), the Joint Committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education referred to these methods of defining what a test measures:

> A psychological construct is . . . a theoretical idea developed to explain and to organize some

> aspects of existing knowledge. . . . It is a dimension understood or inferred from its network of interrelationships. (p. 776)

When hypotheses about people's behavior measured against this construct are proven false experimentally, either the construct definition or the instrument measuring it must be revised (p. 777). This is the stage of research at which the field of reading finds itself today with regard to comprehension.

One major area of argument in defining comprehension is the question of whether it can be operationally defined. Anastasi (1966) insisted that

> the theoretical construct or trait assessed by any test can be defined in terms of the operations performed in establishing the validity of the test. Such a definition should take into account the various external criteria with which the test correlated significantly, as well as the conditions which affect its scores. (p. 309)

The validity of the concept would then be stated in terms of the validity of the test. Bormuth (1970) took the idea of operational definition to the level of the test item. He suggested an operational method of item construction by which he meant that the individual statements of the teacher or text would be manipulated in clearly definable ways to produce the test item (pp. 9-15).

Bormuth's (1970) belief that the syntactic structure of the sentences of instruction can be operationally manipulated is part of his overall view of the communication in a classroom: "Necessary to all uses of achievement

tests is the assumption that there is an abstract structure underlying instruction" (p. 57). These structures may be relationships in the language of the instruction or larger relationships that govern the selection of what is taught: (a) content structure, (b) psychological rules that structure learning, (c) value rules that assign importance to content, and (d) redundancy rules which allow transfer from something taught as an example (pp. 56-65).

Bormuth was explicit about how test items would be formed. He considered this clarity to be a major advance over the "traditional," introspective way of generating test questions. However, his sweeping attacks on conventional tests and test items and his strict reliance on a verbatim transcript of the instruction earned him severe reviews. The difficulty of bringing objectivity to a task which must still produce "intuitively meaningful" results led both Mehrens (1970) and Diederich (1970) to see limitations in Bormuth's approach. They suggested that operational item writing was more suited to the evaluation of specific instructional units with a given content than to the testing of general achievement and skills.

However, Follman (1971) praised the book as a theoretical framework which is "rational, clear, and heuristic enough to serve as a model for achievement test item writing" (p. 57). The difficulty appeared to be in

defining the universe of knowledge and choosing from that what would be considered the significant facts and concepts to be tested. Bormuth's critics would ask immediately: Who is permitted to choose which facts and concepts are significant? What criteria frames the choice? Perhaps the value of his work is that "Bormuth has identified some of the shortcomings of present item-writing techniques" (Mehrens, 1970, p. 667) and opened another approach to stronger test validity.

Bormuth's work is most difficult to apply to the inferential information in a text. His criticisms have aimed at one of the least developed areas of comprehension and testing theory.

The next section will deal with more conventional ways of looking at tests critically.

Subjective and Objective Test Evaluation

Thorndike's Scale for Measuring the Understanding of Sentences and Paragraphs (1917) represents one of the earliest tools for evaluating tests systematically. Most methods of evaluation can be classified as subjective or objective.

Subjective test evaluation includes any which rests on observer opinion or rating of the individual items. This is true regardless of what statistical manipulations are performed on the tally results from

the observers.

Bloom et al. (1956) suggested one way to classify test exercises: first, know what learning has preceded the testing situation; and secondly, attempt to answer the question noting the processes used in identifying the correct response and also in eliminating each of the distractors. This is an unsatisfactory method in light of both the need to assume previous learning and the reliance on self-conscious tracing of cognitive processes which may be automatic or telescoped through practice.

Albright (1927) conducted a subjective inquiry into the comprehension errors made by college students on reading tests by tabulating 20,000 such responses by type of error. Davis (1972) condensed that detailed list into five broad skill areas:

1. Knowledge of word meanings
2. Ability to grasp detailed facts
3. Ability to weave ideas together and made deductions
4. Ability to follow the syntactical structure of a passage
5. Ability to consider a passage objectively without being overwhelmed by personal experiences and feelings

Two more recent subjective evaluations have been based on the extensive work done by Davis in classifying

skills. In the first, Pyrczak (1974) had two observers sort items into three categories:

1. Those answered explicitly or in paraphrase
2. Those requiring the ability to draw inferences
3. Those requiring other skills

The observers were also asked to describe any faults they felt existed in the items.

The rate of agreement on the categorization was only .63 between the observers. Results of the item criticism deviated even more widely: one observer listed 31 faults; the other, 122. The judges obviously worked from different basic viewpoints about test construction.

Stoker and Kropp (1964) confirmed this in talking with their judges after classification of items in their study showed less agreement than expected. In a later discussion of the 1974 study, they pointed out that "the judges were consistent in their different assumptions" (Kropp, Stoker, & Bashaw, 1966, p. 22).

In a second study based on Davis' work, Pyrczak and Rasmussen (1974) completed an analysis of the nine high school reading tests according to the skills measured by each. Again, observer agreement of .65 was found in sorting items into types based on Davis' model of reading comprehension indicating the highly subjective nature of the categorization. This study found wide differences from test to test in both the number of skills measured and the amount of emphasis on each skill.

In these studies, the classification of items is, of course, not a verification that those types of reasoning are actually or exclusively used by readers to answer the items. The classification is also necessarily limited by the categories set up by the researchers.

Auerbach (1971) performed a more thorough analysis on three levels each of three reading tests. Her study attempted to assess the increasing cognitive demands in upper-level questions. Secondly, she also did readability checks on each level and correlated those results with difficulty scores.

In the readability section, at each successive test level, both the number of words and the ratio of difficult words (i.e., non-Spache or non-Dale-Chall words) increased; but the differences between levels were not uniform. Generally, readability (i.e., stylistic elements) and difficulty scores did not correlate highly. In most cases, those correlations were much lower than .70. Based on the fact that these readability factors did increase, however, Auerbach concluded that the higher levels of the tests may contain more remote concepts and more complex ideas.

The most important finding concerning the task requirements at various levels in each test was that pupils needed a considerable prior store of information to perform successfully on this reading test. Thus, a

difference between such tests and either IQ or content achievement tests was blurred.

The actual test items were found to closely resemble items for other school subjects and to measure many of the same kinds of abilities as those other tests. This raised the questions of whether there is a separate construct measured by reading comprehension tests.

The tests Auerbach examined presented similar kinds of selections but showed considerable difference in the distribution of types of questions about them. Number and types of distractors were similar. Characteristic types of selections and questions did, however, appear on certain levels of the tests. Narratives, for example, were more frequent on elementary forms, while secondary selections tended to be factual or content-oriented.

In general, item difficulty appeared to be related not to the sophistication of thought required but to (a) lack of clarity in the selection, (b) the amount of uncommon or academic information required, and (c) the similarity of meaning between the correct answer choice and the distractors.

Auerbach classified the tasks required by questions as either paraphrase or concept tasks. Concept tasks are not differentiated any further, but a schema of nine types of operations was used to specify what a reader used to

answer a question:

1. Recognition
2. Contextual paraphrase
3. Grammatical paraphrase
4. Semantic paraphrase
5. Definite concepts
6. Probable concepts
7. Language concepts
8. Previous knowledge
9. Picture matching

Applying the schema remained a subjective task; observers designated the operations they felt were applied in each item.

To date, the only objective measure of comprehension test items appears to be those provided by the passage dependency studies. Anderson (1972) explained passage dependency in terms of the test and the teaching. The extent to which "a test item measures comprehension depends upon the relationship of the wording of the test item to the wording of the instruction" (p. 167).

Two formulas exist to measure this relationship: Tuinmann's Passage Dependency Index (PDI) and Pyrczak's Index (I) (Pyrczak, 1972). Tuinmann measured the reader's success in answering questions with the corresponding passages available (i.e., the "passage in" condition) as opposed to answering in the absence of the passages (i.e.,

the "passage out" condition). The relative difference is taken as a statistical measure of passage dependency of the questions.

Pyrczak, however, insisted that this technique did not allow for the relative difficulty of the questions themselves. A question may require a complex cognitive operation even in the presence of the selection, and failure to answer correctly may reflect the difficulty of the operation.

The index which Pyrczak proposed is based on a calculation of the probability of a correct answer occurring in the absence of the passage (i.e., the chance factor) versus the actual number of correct responses during testing under a passage-in condition.

Three factors were suggested as sources of information that contaminate the measure of how well a given passage has been understood: (a) general knowledge, (b) information provided by other questions about the same passage, and (c) clues in the relationship of the keyed choice to the distractors; e.g., it is longer, more general, etc.

Pyrczak conceded that his Index gave only one type of information about the items, and that there are at least four others to consider: (a) item difficulty, (b) item discrimination, (c) types of skills required, and (d) the relationship of each item to the overall test

outline.

Further studies have attempted to duplicate and extend this work. Cromer and Wiener (1974) achieved mixed results in their study, concluding that readers do use the paragraphs given on the tests, but also rely on other input in their reasoning. General ability appeared to be a factor as "good readers" outscored "poor readers" in most categories. The criteria for good and poor readers would influence interpretation of these results.

One attempt to eliminate the contaminating influence of questions on reading test scores has been the development of the cloze test. Simons (1971) built a case for this testing technique developed by Taylor (1953):

> The Cloze test is a better measure of reading comprehension than traditional tests because . . . it is not measuring a student's skill in understanding questions. It is not a memory test because the student can continually re-examine the passage. It also does not appear to be measuring a student's familiarity with the content of the passage. . . . The mechanical procedure for developing the test . . . renders it more objective and less subject to the arbitrary judgments of the test constructor than traditional comprehension tests. In addition, a great deal of research has gone into the Cloze test. (Simons, 1971, p. 347)

However, Simons returned to his central theme: "Obviously, the Cloze test lacks construct validity for the same reason that all tests of reading comprehension process do, that is, the absence of a theory of the process" (p. 348).

If analysis finds so many faults with existing

tests, the question remains: Is it possible to write new tests to fit the emerging theory of comprehension?

Davis (1941) claimed to have done so with the Cooperative Reading Comprehension Tests in 1940. Actually, he proceeded in reverse of the above sequence. He began by using a set of questions which he had written and tested, measuring student response to produce a list of variables. That list of variables was used to classify the test questions into a structure of nine operational skills. It was a complex statistical operation but essentially moved from the separate items to a grouping of skills through factor analysis. Davis aimed for a minimum number of factors to explain the maximum amount of variance in reading. In 1972, extending his original studies, he keyed his operational skills to levels in the Bloom et al. Taxonomy.

Although Davis listed nine skill areas, the statistical weight actually supported only the naming of Vocabulary and Reasoning as principal components of reading ability. The statistical analysis used in this study has been called into question (Wark, 1966). Keeping in mind that the Davis study was not theoretically based but rather a psychometrically derived set of variables, researchers must still conclude that the large factor Reasoning masks a variety of cognitive operations.

The other difficulty in using Davis' model as a

test-writing base is that it was postulated as a model of the functioning of mature readers (Davis, 1941, pp. 22-23). It is necessary to question whether it is appropriate for the construction of tests aimed at acquisition and immature stages of reading development.

Recommendations for writing tests existed before Davis' exhaustive effort to capture theory with statistics. Most were subjective and arbitrary. Ebel (1956) suggested "ideal" percentages of six ascending levels of questions which should appear on a comprehension test: content details, 0%; vocabulary, less than 20%; facts, less than 20%; generalizations, more than 10%; understanding, more than 10%; and application, more than 10%.

Michael (1968) and Scriven (1967) suggest other organizing patterns for this sort of test. Educational Testing Service has published a breakdown of its own method for compiling such instruments called ETS Builds a Test (Katz, 1959). They carefully balance various areas of knowledge and levels of cognitive demands in the final item distribution. Such test construction tools will be extremely valuable if and when they can be shown to be based on a viable theory of comprehension.

Anderson (1972) described procedures for writing test questions of six types which may help to clarify the kinds of thinking the questions demand: The first two item types do not strictly require comprehension to answer

correctly although they often appear on comprehension tests: verbatim questions and transposed verbatim questions.

Anderson argued that the simplest level on which comprehension can be assessed is that of paraphrase.

> In order to answer a question based on a paraphrase, a person has to have comprehended the original sentence, since a paraphrase is related to the original sentence with respect to meaning but unrelated to the shape or the sound of the words. (p. 150)

Therefore, paraphrase and transformed paraphrase questions are included in his list. Anderson noted that, with slightly different labels, these are Bormuth's categories as well.

The additional classes of questions which Anderson suggested were questions formed by substituting particular terms for superordinate terms and questions formed by substituting general terms for specific terms. The difficulty with most operational definitions for the way these questions are formed is that they require the questions to be based on a literal transcript of the instruction or a passage from the text, and this limits the types of questions.

Anderson (1972) felt that questions did not always need to be so directly tied to the word-by-word-transcript. The solution he proposed was to "write items that occupy as much semantic space as possible" (p. 160). This rather vague direction apparently meant that questions designed

to measure the comprehension of concepts and principles should be phrased in language and refer to instances of the concept which are different from the ones used in the instruction but which demonstrate the concept. In this case, it would be the concept and not the language of the instruction which would be tested.

Anderson also reviewed test-writing procedures which help decide what to include on a test: (a) a table of specifications, (b) behavioral objectives, (c) a "universe" or "domain" of knowledge, and (d) a random sampling approach to the whole content. The table of specifications has been a traditional tool to balance content and levels of thinking. Anderson found it largely unsatisfactory, but a "reasonable intermediate step in test construction" (p. 160).

Reliance on behavioral objectives implies that the rationale for the inclusion of the objective in the curriculum is accepted by the item writer. There is some value to this approach if that criterion is met, according to Anderson.

Employing the concept of a large universe of possible test items should "insure that at least some fraction of the items entail paraphrase, new instances, or new general terms" (p. 161). Therefore, this method should generate a valid test of the subject. However, the ability to define the universe of possible items

depends on a great deal of knowledge about the semantic structure of discourse which is not available yet. As Anderson points out, this work is in progress.

In the meanwhile, one approach being used in the limited domain of the textbook is the cloze procedure. It was cited as an adaptation of the fourth way to select test items; i.e., "to sample elements from the text or instructional transcript in a partly random fashion" (p. 162).

Each of these methods appears to be partially satisfactory and to be tending in the direction of greater specificity of test content.

In regard to other psychometric methods recommended to improve validity, Anderson stated bluntly: "Manipulating tests to control difficulty level and discriminating power tortures validity" (p. 165).

Finally, he found that both direct information on validity and on the test-writing procedures which affect it are often not reported clearly or completely in technical manuals. Methods of construction and discussion of tests in educational research itself were also found to be disturbingly muddled.

This review of the literature on comprehension and its testing has shown that the field will continue to remain muddled until the relationship between the test and the instruction, and between the test and a

satisfactory theory of comprehension is better understood.

CHAPTER III

PROCEDURES

This study will examine the relationship of Barrett's Taxonomy of Reading Comprehension to seven major tests of reading comprehension. The categories of the taxonomy will be used to categorize items on those instruments. A second judge will verify these placements by categorizing a sample of the items.

An analysis of the results will attempt to answer these questions:

1. What differences will the classification results reveal in the distribution of question types among the seven tests?

2. How frequently will a judge be unable to place an item in one of the available taxonomy categories because the taxonomy is too general, too specific, or inappropriate in the opinion of the judge?

3. How frequently will two judges make identical placements of items in the taxonomy categories?

General Procedure

A sample of reading tests was chosen to be evaluated in terms of the types of questions they contain.

Five taxonomies were reviewed, and one was chosen according to which the test items would be classified. Guidelines for the placement of an item in each category of the taxonomy were prepared.

The 367 items from the seven tests of reading comprehension were classified by the author. An additional category, labeled "problem items," was provided for items which resisted placement. Category frequencies were compared among the seven tests and within each test.

A second judge received a 10% random sample of the population of items which was stratified according to the seven tests from which it was drawn. He classified those 35 items, and the results of that categorization were compared to the results obtained by the first judge.

Data on the number and types of items placed in the problem item category was summarized. Reasons for these placements were examined in an effort to identify consistent difficulties in categorization of items.

Selection of Tests

The seven tests examined in this study were chosen from the eight tests in the Anchor Test Study (1974). Information about the specific forms and levels used is given in Table 2.

One criterion for the inclusion of these tests in the Anchor Study was their widespread use in school testing programs. For that same reason, it was felt that this report of the distribution of question types would be valuable in assessing the role of the tests in those programs.

Description of Tests

General and technical information about the seven tests examined in this study is provided in Tables 2, 3, 4, and 5.

The seven tests of reading comprehension in this study are subtests of major standardized batteries. (See Table 2.) For each, the level designed for the intermediate grades (approximately grades 4 to 6) was chosen. Only the items on the subtest labeled Reading or Reading Comprehension were examined.

These subtests varied in length from 30 to 74 items (see Table 3). The time limits ranged from 25 to 55 minutes. The time per item was not consistent across the tests. The SAT and the MAT placed the most time

TABLE 2

GENERAL DESCRIPTION OF SEVEN STANDARDIZED READING TESTS

Title	Publisher	Edition	Level	Grade Level	Form
California Achievement Tests, 1970 Edition (CAT)	CTB/McGraw-Hill	1970	3	4-6	A
Comprehensive Tests of Basic Skills, Expanded Edition (CTBS)	CTB/McGraw-Hill	1970	2	4.5-6.9	S
Iowa Tests of Basic Skills, Levels Edition (ITBS)	Houghton Mifflin	1971	11	5	6
Metropolitan Achievement Tests (MAT)	Harcourt Brace Jovanovich	1970	Intermediate	5.0-6.9	F
SRA Achievement Tests (SRA)	Science Research Associates	1971	Multilevel Blue	4-5	E
Sequential Tests of Educational Progress, Series II (STEP)	Educational Testing Services	1971	-	4-6	4A
Stanford Achievement Tests (SAT)	Harcourt Brace Jovanovich	1975	Intermediate II	5.0-6.9	A

TABLE 3

TECHNICAL DESCRIPTION OF SEVEN STANDARDIZED READING TESTS

Test	Time Limit (mins.)	Number of Items	Time per Item (mins.)	Reliability		
				KR-20	Spearman-Brown	Parallel Forms
CAT, Section 2 Reading Comprehension	35	42	.83	.87[a] .89[b] .91[c]		
CTBS, Test 2 Reading Comprehension	35	45	.78	.92		
ITBS, Test R Reading Comprehension	55	74	.74		.93	.86
MAT, Test 2, Reading	25	45	.56	.93	.93	
SRA, Reading	50	60	.83	.89[a] .91[b]		
STEP, Reading II	30	30	1.0	.91[a] .92[b] .88[c]		.76[a] .89[b] .88[c]
SAT, Test 2 Reading Comprehension	35	71	.49	.94[b]	.95	

[a]Grade 4.

[b]Grade 5.

[c]Grade 6.

pressure on test takers (.49 and .53 minutes per item, respectively), while the STEP allowed a full minute to complete each item. The other four tests averaged .80 minutes per item.

The difference in time allotment per item did not appear to affect reliability on these instruments. Reliability, which was reported according to three formulas, fell consistently between .87 and .95 with the exception of a parallel forms report of .76 on the Grade 4 level of the STEP. These are average or above average reliabilities for standardized tests available today.

The test format differed widely across the instruments (see Table 4). The majority were a mixture of prose and poetry selections, although the CAT, MAT, and SRA contained no poetry. Only the CAT tested reference skills (i.e., use of a table of contents, an index, and a chart) as part of the Reading Comprehension section. The CTBS, ITBS, SRA, and SAT evaluated these skills in separate sections. This aspect of reading was not included on the MAT or STEP.

Passages on science and social studies topics predominated, with many general narrative selections (see Table 5). The SRA included a fairy tale, the STEP used a short scene from a drama, and the SAT added a biographical sketch. The SAT had the widest variety of

TABLE 4

FORMAT OF TESTS AND ITEMS FOR SEVEN STANDARDIZED READING TESTS

Test	Test Format					Item Format[a]		
	Prose[b]		Poetry	Reference		Sentence-Stem Completion	Question	Imperative
	f	Subjects		f	Subjects			
CAT	3	Science, Social Studies	-	3	Table of Contents, Index, Chart	31	4	-
CTBS	5	Science, Social Studies	2	-	-	19	26	-
ITBS	7	Science, Social Studies, Narrative	2	-	-	2	72	-
MAT	8	Science, Social Studies, Narrative	-	-	-	44	-	1
SRA	6	Science, Social Studies, Narrative, Fairy Tale	-	-	-	33	27	-
STEP II	4	Science, Social Studies, Literature, Drama	1	-	-	14	16	-
SAT	11	Science, Social Studies, Narrative, Biography, Sports	1	-	-	60	11	-

[a]Four possible answers.

[b]Number of paragraphs in each subject varies.

TABLE 5

PUBLISHERS' STATEMENTS OF SKILLS MEASURED ON SEVEN STANDARDIZED READING TESTS

Test	Skills
CAT	1) Understand the meaning of content material 2) Apply rules, facts, concepts, conventions, and principles of problem solving 3) Use tools of reading in progressively difficult situations
CTBS	1) Recognize or recall information 2) Translate concepts from one kind of language (verbal or symbolic) to another 3) Comprehend concepts and their interrelationships 4) Apply techniques, including performing fundamental operations 5) Extend interpretations beyond stated information
ITBS	1) Recognize and understand stated or implied factual details and relationships 2) Discern the purpose or main idea of paragraph 3) Organize ideas 4) Evaluate what is read
MAT	1) Comprehend literally stated material 2) Draw inferences 3) Identify main idea 4) Determine word meaning from context
SRA	1) Restate material 2) Sequence and summarize 3) Draw inferences 4) Apply knowledge to new situations 5) Recognize logical relationships
STEP	1) Comprehend 2) Translate and infer 3) Analyze
SAT	1) Comprehend global meaning 2) Comprehend detailed information 3) Comprehend implied meaning 4) Use context for word and paragraph meaning 5) Draw inferences

types of prose, but it also had the largest number of selections among the seven instruments.

The items were all multiple-choice style, presenting an incomplete stem or a free-standing question followed by four options for the answer. One item on the MAT was a direction to choose an answer and was stated as an imperative.

The relative proportion of the item styles varied widely across the tests but did not seem to be related to the level of reasoning skill required by each item. The question of item style may become more important as work on semantic analysis of tests progresses.

The seven tests also varied in the lists of skills which their publishers claim they test (see Table 5). The lists are basically similar from test to test, but they use many overlapping terms which refer highly related aspects of comprehension or categories which divide the global skill of reading in different ways.

These skills may be as general as to "understand the meaning of content material (CAT)," as limited as "to discern the purpose or main idea of a paragraph (ITBS)," or as specific as to "translate concepts from one kind of language (verbal or symbolic) to another (CTBS)." Only the CTBS, ITBS, and SRA mentioned the higher level skill of evaluation or interpretation as a specific part of the domain of skills they aim to test.

Selection of Taxonomy

Five taxonomies were reviewed and evaluated to select one for use in this study: Barrett (1968, 1972), Bloom et al. (1956), Guszak (1967), Sanders (1966), and Santos (1976). Details of this review are presented in Chapter II and information about the taxonomies summarized in Table 1. For complete texts of the taxonomies, see Appendices A through F.

Based on this review, Barrett's Taxonomy of Reading Comprehension (1972) was chosen as the instrument for categorizing the population of test items.

Description of Taxonomy

The full text of Barrett's Taxonomy of Reading Comprehension (1972), including sample tasks at each level, appears as Appendix D. This taxonomy consists of four major categories of comprehension abilities. Barrett specified that the listed examples are not the only tasks which could require or demonstrate these abilities.

A discussion of the exact relationships among the levels of the taxonomy has been included in Chapter II. Basically, Barrett cautioned that the categories are not perfectly discrete; i.e., a comprehension task may require responses from several categories.

Secondly, he stressed that the categories are not strictly hierarchical; i.e.,

> The learning tasks for any one category can range from easy to difficult and that the difficult tasks from one category may be more intellectually demanding than the easy tasks from another category regardless of the apparent hierarchical ordering of the categories within the Taxonomy. (p. 58)

An outline of the Barrett (1972) categories is given below in his Taxonomy of Reading Comprehension:

1.0 Literal Recognition or Recall

- 1.1 Recognition or Recall of Details
- 1.2 Recognition or Recall of Main Ideas
- 1.3 Recognition or Recall of Sequence
- 1.4 Recognition or Recall of Comparisons
- 1.5 Recognition or Recall of Cause and Effect Relationships
- 1.6 Recognition or Recall of Character Traits

2.0 Inference

- 2.1 Inferring Supporting Details
- 2.2 Inferring the Main Idea
- 2.3 Inferring Sequence
- 2.4 Inferring Comparisons
- 2.5 Inferring Cause and Effect Relationships
- 2.6 Inferring Character Traits
- 2.7 Predicting Outcomes
- 2.8 Inferring about Figurative Language

3.0 Evaluation

- 3.1 Judgments of Reality or Fantasy
- 3.2 Judgments of Fact or Opinion

3.3 Judgments of Adequacy or Validity

3.4 Judgments of Appropriateness

3.5 Judgments of Worth, Desirability, or Acceptability

4.0 Appreciation

4.1 Emotional Response to Plot or Theme

4.2 Identification with Characters or Incidents

4.3 Imagery

The following are criteria for the placement of an item at each of the four major levels. These criteria are based on those provided by Barrett in his full statement of the Taxonomy (see Appendix E). The reader's task for each category is:

1.0 Literal Recognition or Recall

1.1 Recognition--locate or identify explicit statements in the reading selection.

1.2 Recall--produce from memory explicit statements from a selection.

2.0 Inference--use a synthesis of the literal content of a selection, personal knowledge, intuition, and imagination as a basis for conjectures or hypotheses. (A rationale may or may not be required.)

3.0 Evaluation--make judgments about the content of a reading selection by comparing it with external or internal criteria. These may include information from an instructor, an expert, written sources, or

personal experience, knowledge, or values.

4.0 Appreciation--recognize and respond to the literary techniques, forms, styles, and structures used by the author to achieve an emotional as well as an intellectual impact.

Application of Taxonomy

To apply the taxonomy, this writer first read the criteria for placement of an item at each taxonomic level (see Description of Taxonomy and Appendix E). Each of the 367 items on the seven tests was then read and classified using Worksheet A (1 to 7) to record the results. The number of each problem item and the reason for difficulty in placement was recorded (see Appendix H).

To prepare a stratified random sample of items for the second judge, this writer chose an arbitrary sample size of 10% of the original population of items. The sample for Judge B, therefore, was 35 items rather than 36 in order to maintain an even distribution of items across the tests. The sample was stratified according to the seven tests being used. The sample, therefore, consisted of five items from each test. Finally, this writer used a table of random numbers to choose the items from each test to appear in the sample.

To provide a check on the ratings by this writer, the second judge, a doctoral student in the Reading

program at Rutgers University, followed a similar procedure. He read the criteria for placement of an item at each taxonomic level (see Description of Taxonomy and Appendix E). Then, he read and classified each of the 35 items in the sample using Worksheet B (1 through 7) to record his results. Finally, the number of each problem item and the reason for difficulty in placement was recorded (see Appendix J).

Classifications by Judge B were summarized in Appendix I.

Treatment of Data

The classifications by this writer will be tabulated and converted into percentages to represent the portion of a test judged to measure each skill. These results will be reported in Tables 6 and 7 and compared across the seven tests.

A frequency tabulation for the items placed in the problem item category by the first judge will also be reported in Table 6.

The number of identical placements by the two judges will be tabulated and converted into a percentage to indicate inter-rater agreement. Table 8 will show the category assignment both in cases of identical and differing placements. Table 9 will show the types of disagreements in categorization by the two judges.

CHAPTER IV

FINDINGS AND DISCUSSION

This chapter will present and discuss the categorization of test items by two judges. Characteristics and implications of the distribution among the four major categories will be discussed within and across the seven tests. Inter-rater agreement and difficulties in placement of items will also be reported and discussed.

In general, the taxonomy structure appeared to be robust enough to accept every item. While difficulties in placement were noted, these usually arose from questions of the boundaries between categories rather than areas beyond the taxonomy.

The first question of this study was: What differences will the classification results reveal in the distribution of question types across the seven tests?

Judge A was the only person who worked with the total item population of the seven tests. The complete item-by-item categorization by Judge A is given on Worksheets A-1 through A-7 in Appendix G. A summary of that categorization is made in Table 6 showing for each test both the number of items placed in each category and the percentage of the total items which those figures represent.

TABLE 6

DISTRIBUTION OF ITEMS BY CATEGORY FOR SEVEN STANDARDIZED READING TESTS

Test	Total Items	Recognition or Recall		Inference		Evaluation		Appreciation		Problem Items[b]	
		f[a]	%	f	%	f	%	f	%	f	%
CAT	42	14	33	20	48	8	19	0	0	0	0
CTBS	45	10	22	30	67	1	2	4	9	3	7
ITBS	74	11	15	55	74	0	0	8	11	8	11
MAT	45	1	2	43	96	0	0	1	2	7	16
SRA	60	4	7	53	88	2	3	1	2	3	5
STEP	30	2	7	27	90	0	0	1	3	2	7
SAT	71	9	13	62	87	0	0	0	0	10	14

[a] f = frequency.

[b] All problem items were finally placed and counted in a taxonomic category.

Only two tests contained items from all four categories: the CTBS and the SRA. These are not the longest tests in the sample, so length of the test does not appear to be related to the scope of the questions.

Throughout the following discussion, Recognition and Recall items will be referred to simply as Recognition; since on these tests, the reading passages remain available to test takers.

The remaining tests except the SAT were comprised of three types of questions. The SAT was divided between Recognition and Inference items; no Evaluation or Appreciation questions were included. The CAT apparently contained no items tapping Appreciation; while the ITBS, MAT, and STEP each had at least one Appreciation question but no Evaluation items.

The largest category both absolutely and relatively was Inference followed by Recognition. Generally, Evaluation was the smallest group of items on each test with Appreciation also being a rarely included category.

The low percentages in the two upper levels came from very few or, often, one item. The 19% Evaluation set on the CAT and the 11% Appreciation group on the ITBS result primarily from a series of questions on a single atypical selection: a chart in the first case and a poem in the second. When so few item appear in these categories, revising a single placement could drastically

alter the percentages, so caution must be used in interpreting the results. Generally, however, the seven tests appeared to contain few exercises demanding higher-level comprehension processes.

On every test, the Inference category is a larger percentage of the total than Recognition (see Table 7). That margin ranges from 15% (CAT) to 94% (MAT) with a mean of 65% and a median of 74% difference. The Recognition count was as small as 2% (MAT) or 7% (STEP and SRA) of the total, while Inference formed up to 90% (STEP) or 96% (MAT). The CAT came closest to balancing these two categories with 33% Recognition and 48% Inference.

Several of the tests were rated as having very few items in the Recognition category: MAT (1 item), SRA (4 items), and STEP (2 items). Such low counts in Recognition and the general trend toward Inference probably resulted in part from the strict criteria used by Judge A in placing items into Recognition. The difficulty focused on the meaning of the word explicit in the Barrett (1972) description of a Recognition question:

> Recognition Tasks . . . require the student to locate or identify explicit statements in the reading selection itself or in exercises that use the explicit content of the reading selection. Recall Tasks demand the student to produce from memory explicit statements from a selection. (p. 53)

For the purposes of categorization, this judge defined explicit as "word for word" so that substitution

TABLE 7

PERCENTAGE OF RECOGNITION AND INFERENCE ITEMS
ON SEVEN STANDARDIZED READING TESTS

Test	Recognition	Inference	Range[a]
CAT	33	48	15
CTBS	22	67	45
ITBS	15	74	49
MAT	2	96	94
SRA	7	88	81
STEP	7	90	83
SAT	13	87	74

[a] $\underline{M}$ = 65; $\underline{Mdn}$ = 74.

of a synonym or a pronoun made the statement less than explicit. In setting the requirement, it was realized that many items which in ordinary classroom usage and in test writers' manuals would be considered as factual, would now be eliminated from the Recognition classification. In turn, the Inference category would receive many items of a very basic level of translation which may be considered inferential to some degree.

Admittedly, there is a wide gap between the levels of inferential thought required by the following hypothetical items:

1. Jane praised her sister's accomplishment. What did Jane do?
   (a) ignored her sister
   (b) paid her sister a compliment
   (c) talked to her sister
2. Bill went to the store. He bought ice cream for his family. Who bought ice cream?
   (a) Bill
   (b) his family
   (c) a neighbor
3. Mary came in the door, shook her dripping umbrella, and took off her wet coat. What was the weather that day?
   (a) sunny
   (b) windy

(c) rainy

However, in order to keep the Recognition category as "clean" as possible, all three of these would be considered to involve Inference.

Of all the tests, the CAT showed the greatest reliance on phrases taken verbatim from the selection as answers. To a lesser degree, the CTBS, ITBS, and the SAT also exhibited this tendency. The MAT, SRA, and STEP phrased most of their questions so that some form of inference would be required.

Because of the limitations of the data discussed previously, it is doubtful whether the appearance of other categories can be used as an indicator of a broader scope for any of the tests. Finer judgments about the relative merits of the tests would require a better understanding of the cognitive demands of each test. This would necessitate the development of a scale for ranking questions in the Inference category. Relative difficulty of the types of inference employed would have to be demonstrated.

As Spache pointed out, this had not yet been done in 1963; in 1975, Panel Report 1 from the NIE Conference on Studies in Reading listed inference as a prime area for further study (Miller, 1976, p. 721).

The second question of this study was: How frequently will a judge be unable to place an item in one

of the available taxonomy categories because the category is too general, too specific, or inappropriate in the opinion of the judge?

Judge A marked 33, or 9%, of the 367 items as "problem items." However, the question about each of these items was resolved; and all 33 were included finally in the count for one of the four categories.

The difficulties were of three types:

1. Many of the tests included items asking for a choice of a "good" or "the best" title for a selection. The word "best" implied the use of some internal criteria on the part of the reader. However, the purpose of these questions seemed to this judge to be simpler than Evaluation which Barrett (1972) defined as making

> judgments about the content of [the] reading, judgments that have to do with it accuracy, acceptability, worth, desirability, completeness, suitability, timeliness, quality, truthfulness, or probability of occurrence. (p. 55)

Use of the word "best" appeared to be simply an artifact of the way the question is phrased. Barrett listed Recognition or Recall of Main Ideas as Category 1.2 and Inferring the Main Idea as Category 2.2 (pp. 53-54).

2. Several questions asked for meanings of vocabulary words from the passage. These could be Recall tasks if the meanings were part of the student's prior knowledge or Inference tasks if the words were unknown or capable of several meanings depending on the context. In most

cases, Judge A placed these in the Inference category.

3. Many of the questions involving pronoun reference to determine the answer came very close to placement as Recognition items. The effect of simple and complex sentence structure makes some pronoun references more obscure than others. This made it difficult to follow a single rule for such items. However, a strict interpretation was followed, moving these items to Inference.

Judge B also met some dilemmas in classification, particularly the second and third difficulties mentioned; but all of the items in the sample were eventually placed into categories. His discussion of specific problems appears in Appendix J. They involve 8 items or 22.8% of his sample.

Judge B offered the following rationale for his own distinction between Recognition and Inference placements for vocabulary items:

1. Questions requiring slight translations of simple words are classed as literal (Recognition). Although students are not called upon to recall or recognize explicit statements, they do not have to conjecture or intuit either.

2. Questions that seem to ask for word meaning in context are classed as Inference. Assuming the word is unfamiliar, the students must synthesize its meaning.

The third question of this study was: How frequently will two judges make identical placements of items in the taxonomy categories?

Categorizations of the 35 items in the sample by both judges are given in Table 8.

Sixteen items in the sample, or 45.7%, were categorized in the same way by both judges. Thirteen of these placements, or 81%, of the agreements were Inference, while three, or 19%, were Recognition.

There was no agreement on items involving placements by one or the other judge in the Evaluation or Appreciation category. This may be related to the fact that this sample formed a small percentage of the total number of items; and according to Judge A's results, few of these higher-level questions occur even in that population.

The rate of agreement in this study (45.7%) is lower than that obtained in most previous studies of this type. Pyrczak and Rasmussen (1974), for example, achieved a 65.4% rate of agreement in placement. Teams of raters in the Stoker and Kropp study (1966) showed unanimous decisions on 31% of the items on one paragraph and for 28% of the items on another. Varying numbers of categories and raters may make these studies difficult to equate, however.

In the present study, 19 items in the sample, or

TABLE 8

CATEGORIZATION OF SAMPLE OF ITEMS FROM SEVEN STANDARDIZED READING TESTS BY TWO JUDGES

Test	Item f=35	Agreements f=16 (45.7%)	Disagreements f=19 (54.3%)	
		Judges A & B	Judge A	Judge B
CAT	9	Recognition		
	12		Inference	Recognition
	16		Inference	Recognition
	24		Inference	Recognition
	38		Evaluation	Recognition
CTBS	9	Inference		
	14	Recognition		
	31	Inference		
	40		Recognition	Inference
	45		Recognition	Inference
ITBS	25		Recognition	Inference
	47		Inference	Recognition
	58		Inference	Evaluation
	72		Inference	Evaluation
	87	Inference		
MAT	1		Inference	Recognition
	12		Inference	Recognition
	23	Inference		
	33	Inference		
	36	Inference		
SRA	2	Inference		
	25		Inference	Recognition
	29		Inference	Evaluation
	30	Inference		
	48	Inference		
STEP	7		Inference	Recognition
	17		Inference	Appreciation
	28	Inference		
	29	Inference		
	30		Inference	Evaluation
SAT	3	Recognition		
	8		Inference	Recognition
	36	Inference		
	39		Inference	Recognition
	64	Inference		

54.3%, were categorized differently by the two judges. The types of disagreements are shown in Table 9. Eighteen (94.7%) involved the use of the Inference category by one or the other judge.

Thirteen involve a choice between Inference and Recognition. In 10 cases, or 56.2% of the total disagreements, Judge A chose Inference and Judge B chose Recognition. In three cases, or 15.8% of the total, the reverse occurred.

In four cases, the items that were rated by Judge A simply as Inference were placed by Judge B in Evaluation. One item was rated as Inference by Judge A but seen as Appreciation by Judge B.

Of the 19 placements which were disputed, 15, or 78.9%, involved Judge A's use of Inference. However, of those 15, only 10, or 52.6%, of the total disagreements placed an Inference rating by Judge A against a Recognition rating by Judge B. Five, or 26.3%, of the total set an Inference rating by Judge A against a placement in Evaluation or Appreciation by Judge B. Since these levels have been defined as including an inference component, these five may be seen as less serious disagreements.

Clearly, the problems in identifying item purposes revolved around consensus on the definition and scope of the inference process. The strict interpretation by

TABLE 9

TYPES OF DISAGREEMENTS IN CATEGORIZATION OF SAMPLE OF ITEMS FROM SEVEN STANDARDIZED READING TESTS BY TWO JUDGES

Disagreement		f	Percent[a]
Judge A	Judge B		
Ratings Involving Inference[b]			
Inference[b]	Recognition	10	52.6
Recognition	Inference	3	15.8
Inference[b]	Evaluation	4	20.8
Inference[b]	Appreciation	1	5.2
Total		18	94.7
Ratings Not Involving Inference			
Evaluation	Recognition	1	5.2
TOTAL DISAGREEMENTS		19	99.9[b]

[a]Rounded to nearest tenth.

[b]Fifteen (78.9%) of the total disagreements involved Inference ratings from Judge A and non-Inference ratings from Judge B.

Judge A of the differences between Recognition and Inference did not account for as many of the differences in rating as expected.

CHAPTER V

SUMMARY AND CONCLUSIONS

This chapter summarizes the study and presents conclusions regarding the four questions considered. Suggestions are also included for further research.

Summary

The purpose of this study was to investigate the relationship of seven reading comprehension tests to the construct of comprehension which they claim to measure. This relationship was investigated by the subjective rating of items on those tests by two judges according to a major taxonomy of reading comprehension.

Seven subtests of reading comprehension were chosen from the group of tests in the Anchor Test Study. The most recent edition of each test was used: California Achievement Test, Form A (1970); Comprehensive Tests of Basic Skills, Expanded Edition, Form S (1963-1970); Iowa Tests of Basic Skills, Form 6 (1971); Metropolitan Achievement Tests, Intermediate, Form F (1970); SRA Achievement Series, Blue Level, Form E (1971); Sequential Tests of Educational Progress, Series II, Form 4A (1969); and Stanford Achievement Test, Intermediate Level II, Form A

(1973).

Five taxonomies were reviewed, and Barrett's Taxonomy of Reading Comprehension (1972) was chosen on the basis of its applicability to pencil-and-paper tests and its clear-cut categories. This writer classified 367 items according to Barrett's criteria into four types of questions: Recognition and Recall, Inference, Evaluation, and Appreciation. The percentage of items in each category was compared within and across tests.

A second judge classified 35 randomly selected items in a sample stratified according to the seven tests. Results of that classification were compared to placements by the first judge.

Reasons for the difficulties experienced by both judges in placing items were collected and examined in order to isolate weaknesses in the placement criteria, the taxonomy, the items, and the definitions used in discussing comprehension.

Conclusions

The categorization tally for both judges indicated that Inference items are the most common on each of these seven tests. Recognition tasks rank next in frequency. These two categories account for the overwhelming majority of the test items, while those tapping Evaluation and Appreciation are much less developed. Evaluation items

appear only slightly more often than Appreciation questions on most of the instruments.

The relative weight of Recognition and Inference items varies across the tests. Judge A, in particular, found that questions requiring explicit word-for-word responses from the text are not as prevalent as has been claimed. At the same time, a wider range of items were found to involve Inference than expected. The comprehensiveness of the category was partially due to its definition in this study. The processes involved in this category need to be clarified, particularly the role of Inference as a common element of each category.

The lack of agreement in the research and in subjective ratings for Inference and higher-level tasks raises the question of whether there are actually clear-cut differences among these levels. Perhaps what we are dealing with are different applications of the basic inference process. Rather than a cumulative progression in the three processes (Inference, Evaluation, and Appreciation), there may be a lateral relationship in which some aspects of processing are shared by the three; but none is completely subsumed in another. In order to clarify this relationship, a better understanding of the very simplest level of inferential translation needs to be developed. From there, fuller descriptions of the inference process at many levels of sophistication can

be attempted.

The finding that most of the tests contained few questions designated as related to "higher level" processes should not be construed as "blaming" the test writers for the distribution of items. The analysis is a descriptive one. It may be that the highest levels are not critical on a test of comprehension but that the many ways in which students exercise inference do need to be evaluated.

The second conclusion of this study was that the judges encountered varying amounts of difficulty in categorization. The uncertainty was partially a matter of definition. Different basic assumptions about what characterizes inference clearly affected each person's judgments. The phrasing of many of the items required the judges to refine their definitions of the categories during item placement.

Difficulty in placement also occurred because the test writers did not set out in most cases to construct items to fit specifically into single categories of the Taxonomy. This categorization came as a post-facto effort, and discrepancies are to be expected.

Difficulties centered on items (a) asking for the "best" answer, (b) asking for the meaning of a vocabulary term, and (c) involving pronoun reference in the text or the answer choices.

The third conclusion was that the low rate of

inter-judge agreement in item categorization was a further demonstration of the problems inherent in this type of analysis. Placements by the two judges were identical for 46% of the items of the sample. It had been anticipated that the smaller number of categories used in this study might generate more consensus than other studies which used, for example, five or seven categories had been able to achieve. This expectation was not fulfilled.

The design of these studies cannot control for the subjective interpretation each judge places on the criteria, the passages, the questions, and the response options. The effect, the range of subjective ratings for each item is a result of the operational definitions of the categories which each judge formed.

In summary, the results in a study of this type depend on two variables chosen by the investigator: (a) the persons who perform the categorization, and (b) the labels used for the process of comprehension. Different results would occur with other judges because the placements are subjective and open to dispute. The results on the same items would also change if the process were viewed through a different taxonomy or other skill list.

Therefore, more explicit criteria should be devised to use in these analyses if they are to yield more useful information. Also, a comparison of analyses on the same set of items according to different organizing schemes

may help to find the common ground among the schemes and contribute to an understanding of the process they represent.

To avoid endless documentation of individual differences in identifying the intent of items, new methods of linking the question with the process it evokes must be found. The definitions need to be more closely tied to the levels of the Taxonomy or to a scheme which is more comprehensive than the linear hierarchy.

The final conclusion is that the dilemma and the discrepancies of this attempt at taxonomic categorization may center on this dichotomy between a linear taxonomy and a holistic thought process which results in comprehension. Present tools are unable to depict a system which uses both linear and holistic thought, but comprehension appears to involve both kinds of processing. Future research needs to investigate and specify the dual nature of thought in language processes.

Considerations in Taxonomic Analysis

The difficulty in using the taxonomic format to represent both linear and holistic thought is the need to indicate the dimension of time in the thinking process. In taxonomic classification, an item is labeled as evoking a single process; while, in actuality, the item may require the use of several processes in a flexible time

sequence.

A taxonomy depicts essentially a single channel treatment of information. In describing comprehension, it cannot handle the possibility of parallel processing of information, nor can it clearly show the dynamics of the interaction between the results of such simultaneous compiling of information. Therefore, as a tool for the analysis of language processes, taxonomies are critically limited.

In order to hypothesize ways in which the higher-level thought processes interact, the sequential organization of the taxonomy categories may need to be reconceptualized. For example, the components of the general process which Barrett designated as inferential thought might be described in terms which cut across or combine Translation, Application, Analysis, and Synthesis in a different time frame. In such a system, the higher processes may precede and organize the simpler ones, while at the same time relying upon the results of the basic processes for the raw material with which to work.

Therefore, a single item cannot only evoke a single process since comprehension by definition would be operating in a dynamic situation that is unlike a two-dimensional taxonomy.

REFERENCES

Albright, B. F. Typical reading disabilities of college entrants. Unpublished master's thesis, University of Southern California, 1927.

Alshan, L. M. A factor-analytic study of the items used in the measurement of some fundamental factors of reading comprehension. Unpublished doctoral dissertation, Teachers College, Columbia University, 1964.

Anastasi, A. Some current developments in the measurement and interpretation of test validity. In A. Anastasi (Ed.), Testing problems in perspective. Washington, D.C.: American Council on Education, 1966.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Auerbach, I. An analysis of reading comprehension tests: Final report. Project No. 0-A-074. Grant No. 0EG-1-71-0001. Cambridge, Mass.: Harvard University Graduate School of Education, 1971.

Ayers, J. D. Justification of Bloom's taxonomy by factor analysis. Paper read at the annual meeting of the American Education Reading Association, Chicago, February, 1966.

Barrett, T. C. The Barrett taxonomy: Cognitive and affective dimensions of reading comprehension. Cited in T. Clymer, What is reading?: Some current concepts. In H. M. Robinson (Ed.), Innovation and change in reading instruction. The sixty-seventh yearbook of the National Society for the Study of Education, part II. Chicago: National Society for the Study of Education, 1968, pp. 19-23.

Barrett, T. C. Taxonomy of reading comprehension. Cited in R. J. Smith & T. C. Barrett, Teaching reading in the middle grades. Reading, Mass.: Addison-Wesley, 1974. [Reprinted from T. C. Barrett, Taxonomy of reading comprehension, Reading 360 monograph. Lexington, Mass.: Ginn, 1972.]

Berg, P. C. Evaluating reading abilities. In W. H. MacGinitie (Ed.), Assessment problems in reading. Newark, Del.: International Reading Association, 1973.

Betts, E. A. Reading is thinking. The Reading Teacher, 1961, 15, 179-184.

Bianchini, J. C., & Loret, P. G. Anchor test study: Final report. Berkeley, Calif.: Educational Testing Service, 1974.

Bloom, B. S., Engelhart, M., Furst, E., Hill, W., & Krathwohl, D. (Eds.). Taxonomy of educational objectives. New York: McKay, 1956.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Buros, O. K. (Ed.). The seventh mental measurement yearbook. Highland Park, N.J.: Gryphon Press, 1972.

Buros, O. K. (Ed.). Tests in print II: An index to tests, test reviews, and the literature on specific tests. Highland Park, N.J.: Gryphon Press, 1974.

Buros, O. K. (Ed.). Reading tests and reviews, II. Highland Park, N.J.: Gryphon Press, 1975.

Calfee, R. C. Sources of dependency in cognitive processes. In D. Klahr (Ed.), Cognition and instruction. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

California Achievement Tests. Level 3, Form A. Monterey, Calif.: CTB/McGraw-Hill, 1970.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carver, R. P. Toward a theory of reading comprehension and rauding. Reading Research Quarterly, 1977-1978, 13, 8-63.

Cleland, D. L. A construct of comprehension. Reading and Inquiry, IRA Proceedings, 1965, 10, 59-64.

Clymer, T. What is "reading"?: Some current concepts. In H. M. Robinson (Ed.), Innovation and change in reading instruction. The sixty-seventh yearbook of the National Society for the Study of Education, part II. Chicago: National Society for the Study of Education, 1968.

Comprehensive Tests of Basic Skills, Expanded Edition. Level 2, Form S. Monterey, Calif.: CTB/McGraw-Hill, 1970.

Cromer, W., & Wiener, M. Do reading tests measure "reading"? Worchester, Mass.: Clark University, 1974.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Davis, F. B. Fundamental factors of comprehension in reading. Unpublished doctoral dissertation, Harvard University, 1941.

Davis, F. B. (Ed.). The literature of research in reading with emphasis on models. New Brunswick, N.J.: Rutgers University, 1971.

Davis, F. B. Psychometric research on comprehension in reading. Reading Research Quarterly, 1972, 7, 628-678.

Davis Reading Test. New York: Psychological Corporation, 1962.

Diederich, P. B. Review. Educational and Psychological Measurement, 1970, 30, 1003-1005.

Ebel, R. L. Some limitations of criterion-referenced measurement. In G. H. Bracht, K. D. Hopkins, & J. C. Stanley, Perspectives in educational and psychological measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Eller, W., & Farr, R. Consumer awareness in testing reading. In S. F. Wanat (Ed.), Linguistics and reading series: 1. Issues in evaluating reading. Arlington, Va.: Center for Applied Linguistics, 1977.

Farr, R. Reading: What can be measured? Newark, Del.: International Reading Association, 1969.

Farr, R., & Anastasiow, N. Tests of reading readiness and achievement: A review and evaluation. Newark, Del.: International Reading Association, 1969.

Follman, J. Review. Journal of Educational Measurement, 1971, 8, 57-59.

Gagné, R. The acquisition of knowledge. Psychological Review, 69, 355-365.

Gibson, E. J., & Levin, H. The psychology of reading. Cambridge, Mass.: MIT Press, 1975.

Goodman, K. S. Linguistically sound research in reading. In R. Farr, S. Weintraub, & B. Tone (Eds.), Improving reading research. Newark, Del.: International Reading Association, 1976.

Gough, P. B. One second of reading. In J. F. Kavanagh, & I. G. Mattingly (Eds.), Language by ear and by eye: The relationships between speech and reading. Cambridge, Mass.: MIT Press, 1972.

Guszak, F. J. Teacher questioning and reading. The Reading Teacher, 1967, 21, 227-234.

Harker, W. J. An evaluation summary of models of reading comprehension. Journal of Reading Behavior, 1973, 5, 26-34.

Harker, W. J. Classroom implications from models of reading comprehension. In B. S. Schulwitz (Ed.), Teachers, tangibles, techniques: Comprehension of content in reading. Newark, Del.: International Reading Association, 1975.

Hilgard, E., & Marquis, D. Conditioning and learning (2nd ed.). New York: Appleton-Century Crofts, 1961.

Hill, W. R. Concerning reading theory: A reaction to Carver's 'Toward a theory of reading comprehension and rauding.' Reading Research Quarterly, 1977-1978, 13, 64-91.

Holmes, J. A. Factors underlying major reading disabilities at the college level. Unpublished doctoral dissertation, University of California, Berkeley, 1948.

Holmes, J. A. Basic assumptions underlying the substrata-factor theory. Reading Research Quarterly, 1965, 1, 4-28.

Huey, E. B. The psychology and pedagogy of reading. Cambridge, Mass.: MIT Press, 1968. (Originally published, 1908.)

Hunkins, F. P. The influence of analysis and evaluation questions on achievement in sixth grade social studies. Educational Leadership, 1968, 25, 326-332.

Hunt, L. C. Can we measure specific factors associated with reading comprehension? Journal of Educational Research, 1957, 51, 161-171.

Iowa Tests of Basic Skills, Levels Edition. Level 11, Form 6. Boston: Houghton Mifflin, 1971.

James, W. On the function of cognition. Mind, 1928, 37, 1-18.

Jenkinson, M. D. Sources of knowledge for theories of reading. Journal of Reading Behavior, 1969, 1, 11-29.

Katz, M. R. ETS builds a test. Princeton, N.J.: Educational Testing Service, 1959.

Kerfoot, J. F. Problems and research considerations in reading comprehension. The Reading Teacher, 1965, 18, 250-256.

Kerlinger, F. N. Foundations of behavioral research. New York: Holt, Rinehart, & Winston, 1965.

Kingston, A. J. The measurement of reading comprehension. In O. Causey & E. Bliesmer (Eds.), Research and evaluation in college reading. The ninth yearbook of the National Reading Conference for College and Adults. Fort Worth: Texas Christian University Press, 1960.

Klein, M. F. Uses of the taxonomy of educational objectives (cognitive domain) in constructing tests for primary school pupils. The Journal of Experimental Education, 1972, 40, 38-50.

Körner, S. Classification theory. Encyclopaedia Brittannica (15th ed.). Macropaedia: Knowledge in Depth, 4. Chicago: Encyclopaedia Britannica, 1974.

Kropp, R. P., & Stoker, H. W., with the assistance of Bashaw, W. L. The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives. Cooperative Research Project No. 2117. Tallahassee: Institute of Human Learning and Department of Educational Research and Testing, Florida State University, 1966.

Locke, J. Of the conduct of the understanding. In F. W. Garforth (Ed.), John Locke's of the conduct of the understanding. New York: Teachers College Press, 1966. (Originally published, 1697.)

McCollough, C. M. Pioneers of research in reading. In H. Singer & R. B. Ruddell, Theoretical models and processes of reading (2nd ed.). Newark, Del.: International Reading Association, 1976.

Madaus, G. F., Woods, E. M., & Nuttall, R. L. A causal model analysis of Bloom's taxonomy. American Educational Research Journal, 1973, 10, 253-262.

Mavrogenes, N. A., Winkley, C. K., Hanson, E., & Vacca, R. T. Concise guide to standardized secondary and college reading tests. Journal of Reading, 197[illegible], 18, 12-21.

Mehrens, W. A. Review. Contemporary Psychology, 1970, 15, 666-667.

Metropolitan Achievement Tests. Intermediate Level, Form F. New York: Harcourt Brace Jovanovich, 1970.

Michael, J. J. Structure of intellect theory and the validity of achievement examinations. Educational and Psychological Measurement, 1968, 28, 1141-1149.

Miller, G. A. Text comprehension skills and process models of text comprehension. In H. Singer & R. B. Ruddell (Eds.), Theoretical models and processes of reading (2nd ed.). Newark, Del.: International Reading Association, 1976.

Pryczak, F. Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. Reading Research Quarterly, 1972, 8, 62-71.

Pyrczak, F. Subjective evaluation of the quality of standardized reading comprehension items. Washington, D. C.: National Science Foundation, 1974.

Pyrczak, F., & Rasmussen, M. Skills measured by selected reading tests designed for high school use. Reading Improvement, 1974, 11, 5-8.

Roberts, N. Further verification of Bloom's taxonomy. Journal of Experimental Psychology, 1976, 45, 16-19.

Robinson, H. M. (Ed.). Innovation and change in reading instruction. The sixty-seventh yearbook of the National Society for the Study of Education, part II. Chicago: University of Chicago Press, 1968.

Ryor, J. NEA again says: Get rid of standardized tests. NEA Reporter, Oct., 1977, 16(7), 3.

Rystrom, R. Reflections of meaning. Journal of Reading Behavior, 1977, IX, 193-200.

SRA Achievement Series. Multilevel Blue, Form E. Chicago: Science Research Associates, 1971.

Samuels, S. J., Begy, G., & Chen, C. C. Comparison of word recognition speed and strategies of less skilled and more highly skilled readers. Reading Research Quarterly, 1975, 11, 72-86.

Sanders, N. M. Classroom questions: What kinds? New York: Harper & Row, 1966.

Santos, I. A classification scheme for reading questions. Cited in H. W. Forgan & C. T. Mangrum II, Teaching content area reading skills: A modular preservice and inservice program. Columbus, Ohio: Charles E. Merrill, 1976, 169-220.

Schultz, R. E. Testing theories of reading instruction. In J. Bormuth (Ed.), Research designs in reading. Newark, Del.: International Reading Association, 1967.

Scriven, M. The methodology of evaluation. In R. W. tyler, R. M. Gagné, & M. Scriven, Perspectives of curriculum evaluation. American Education Research Association Monograph Series on Curriculum Evaluation No. 1. Skokie, Ill.: Rand McNally, 1967.

Sequential Tests of Educational Progress, Series II. Form 4A. Princeton: Educational Testing Services, 1971.

Simons, H. D. Reading comprehension: The need for a new perspective. Reading Research Quarterly, 1971, 6, 338-363.

Singer, H. Substrata-factor theory of reading: Theoretical design for teaching reading. In H. Singer & R. B. Ruddell (Eds.), Theoretical models and processes of reading (2nd ed.). Newark, Del.: International Reading Association, 1976. [Reprinted from J. A. Figurel (Ed.), Challenge and experiment in reading. Proceedings of the International Reading Association Annual Convention, 1962, 7, 226-232.]

Singer, H. Substrata-factor patterns accompanying development in power of reading, elementary through college level. In H. Singer & R. B. Ruddell (Eds.), Theoretical models and processes of reading (2nd ed.). Newark, Del.: International Reading Association, 1976. [Reprinted from E. Thurston & L. Hafner (Eds.), Philosophical and sociological bases of reading. Fourteenth yearbook of the National Reading Conference, 1964, 41-56.]

Skinner, B. F. The distribution of associated words. Psychological Record, 1937, 1, 69-76.

Skinner, B. F. Are theories of learning necessary? Psychological Review, 1950, 57, 193-216.

Slobin, D. I. Psycholinguistics. Glenview, Ill.: Scott, Foresman, 1971.

Smith, F. Psycholinguistics and reading. New York: Holt, Rinehart & Winston, 1973.

Smith, I. L. I.Q., creativity, and the taxonomy of educational objectives: Cognitive domain. The Journal of Experimental Education, 1970, 38, 58-60.

Smith, M. Questions and reading comprehension. Unpublished paper submitted to Dr. M. Kling, Graduate School of Education, Rutgers--The State University of New Jersey, December, 1977.

Spache, G. D. Toward better reading. Champaign, Ill.: Garrard, 1963.

Stanford Achievement Test. Intermediate II Level, Form A. New York: Harcourt Brace Jovanovich, 1975.

Stanley, J. C., & Bolton, D. L. A review of Bloom's 'Taxonomy of educational objectives' and J. R. Gerberich's 'Specimen objective test items, A guide to achievement test construction.' Educational and Psychological Measurement, 1957, 17, 631-634.

Stanley, J. C., & Hopkins, K. D. Educational and Psychological Measurement and Evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Stauffer, R. G. Teaching reading as a thinking process. New York: Harper & Row, 1969.

Stedman, C. H. An analysis of the assumptions underlying the taxonomy of educational objectives: Cognitive domain. Journal of Research in Science Teaching, 1973, 10, 235-241.

Stennett, R. G., Smythe, P. C., & Hardy, M. Hierarchical organization of reading subskills: Statistical approaches. Journal of Reading Behavior, 1957, 7, 223-228.

Stoker, H. W., & Kropp, R. P. Measurement of cognitive processes. Journal of Educational Measurement, 1964, 1, 39-42.

Taylor, W. L. Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 1953, 30, 414-438.

Thorndike, E. Reading as reasoning: A study of mistakes in paragraph meaning. Journal of Educational Psychology, 1917, 8, 323-332.

Thurstone, L. L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

Tuinmann, J. J. Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 1973-1974, 9, 206-223.

Wanat, S. F. Evaluating reading: An introduction. In S. F. Wanat (Ed.), Linguistics and reading series: 1. Issues in evaluating reading. Arlington, Va.: Center for Applied Linguistics, 1977.

Wardhaugh, R. Reading: A linguistic perspective. New York: Harcourt, Brace & World, 1969.

Wark, D. M. Substrata-factor theory: A dissenting view. In G. B. Schick & M. M. May (Eds.), New frontiers in college-adult reading. Fifteenth yearbook of the National Reading Conference. Milwaukee: National Reading Conference, 1966.

Weaver, W. W. The contribution of research to reading theory. In A. J. Kingston (Ed.), Toward a psychology of reading and language: Selected papers of Wendell W. Weaver. Athens: University of Georgia Press, 1977. (a)

Weaver, W. W. Sources of information for responses to reading test items. In A. J. Kingston (Ed.), Toward a psychology of reading and language: Selected papers of Wendell W. Weaver. Athens: University of Georgia Press, 1977. (b)

Wittgenstein, L. Preliminary studies for the philosophical investigations generally known as the Blue and Brown books. Oxford: Blackwell, 1958.

Wittrock, M. C. The generative model of reading comprehension. Journal of Reading Behavior, 1977, 9, 109-112.

APPENDIX A

CONDENSED VERSION OF THE TAXONOMY OF EDUCATIONAL OBJECTIVES[1]

Cognitive Domain

KNOWLEDGE

1.00 KNOWLEDGE
- 1.10 Knowledge of Specifics
  - 1.11 Knowledge of Terminology
  - 1.12 Knowledge of Specific Facts
- 1.20 Knowledge of Ways and Means of Dealing with Specifics
  - 1.21 Knowledge of Conventions
  - 1.22 Knowledge of Trends and Sequences
  - 1.23 Knowledge of Classifications and Categories
  - 1.24 Knowledge of Criteria
  - 1.25 Knowledge of Methodology
- 1.30 Knowledge of Universals and Abstraction in a Field
  - 1.31 Knowledge of Principles and Generalizations
  - 1.32 Knowledge of Theories and Structures

INTELLECTUAL ABILITIES AND SKILLS

2.00 COMPREHENSION
- 2.10 Translation
- 2.20 Interpretation
- 2.30 Extrapolation

3.00 APPLICATION

4.00 ANALYSIS
- 4.10 Analysis of Elements
- 4.20 Analysis of Relationships
- 4.30 Analysis of Organizational Principles

---

[1] B. S. Bloom, M. Engelhart, E. Furst, W. Hill, & D. Krathwohl. Taxonomy of Educational Objectives (New York: David McKay, 1956), pp. 201-207.

5.00 SYNTHESIS
- 5.10 Production of a Unique Communication
- 5.20 Production of a Plan or Proposed Set of Operations
- 5.30 Derivation of a Set of Abstract Relations

6.00 EVALUATION
- 6.10 Judgments in Terms of Internal Evidence
- 6.20 Judgments in Terms of External Criteria

APPENDIX B

SANDERS' PRELIMINARY DEFINITIONS OF BLOOM'S CATEGORIES OF THINKING[1]

1.	MEMORY	The student recalls or recognizes information.
2.	TRANSLATION	The student changes information into a different symbolic form or language.
3.	INTERPRETATION	The student discovers relationships among facts, generalizations, definitions, values, and skills.
4.	APPLICATION	The student solves a lifelike problem that requires the identification of the issue and the selection and use of appropriate generalizations and skills.
5.	ANALYSIS	The student solves a problem in the light of conscious knowledge of the parts and forms of thinking.
6.	SYNTHESIS	The student solves a problem that requires original, creative thinking.
7.	EVALUATION	The student makes a judgment of good or bad, right or wrong, according to the standards he designates.

[Note.] The reader is reminded that these preliminary definitions are by no means adequate for distinguishing the categories. They are presented only as a necessary background to a discussion of the way in which ideas from the Taxonomy of Educational Objectives can be used by classroom teachers.

---

[1] N. M. Sanders. Classroom Questions: What Kinds? (New York: Harper & Row, 1966), p. 3.

APPENDIX C

READING COMPREHENSION QUESTION-RESPONSE INVENTORY[1]

RECOGNITION. These questions call upon the students to utilize their literal comprehension skills in the task of locating information from reading context. Frequently, such questions are employed in the guided reading portion of a story, i.e., Find what Little Red Ridinghood says to the wolf.

RECALL. Recall questions, like recognition questions, concern literal comprehension. The recall question calls upon the student to demonstrate his comprehension by recalling factual material previously read. Generally, such activity is primarily concerned with the retrieval of small pieces of factual material, but the size of the pieces may vary greatly. An example of a recall question would be the following where the answer to the question is clearly printed in the text, i.e., What was Little Red Ridinghood carrying in the basket?

TRANSLATION. Translation questions require the student to render an objective, part-for-part rendering of a communication. As such, the behavior is characterized by literal understanding in that the translator does not have to discover intricate relationships, implications, or subtle meanings. Translation questions frequently call upon students to change words, ideas, and pictures into different symbolic form as is illustrated in the following from Bloom (1956).

Translation from one level of abstraction to another; abstract to concrete, lengthy to brief communication, i.e., Briefly re-tell the story of Little Red Ridinghood.

Translation from one symbolic form to another, i.e., Draw a picture of the first meeting between Little Red Ridinghood and the wolf.

Translation from one verbal form to another; non-literal statements to ordinary English (metaphor, symbolism).

---

[1]F. J. Guszak. Teacher questioning and reading. The Reading Teacher, 1967, 21, pp. 228-231.

CONJECTURE. These questions call for a "cognitive leap" on the part of the student as to what will happen or what might happen. As such, the conjecture is an anticipatory thought and not a rationale, i.e., What do you think that Little Red Ridinghood will do in the future when she meets a wolf in the forest?

EXPLANATION. Explanation questions, like conjecture questions, are inferential in nature. However, the inference involved in the explanation situations calls upon the student to supply a rationale. The rationale must be inferred by the student from the context developed or go beyond it if the situation is data-poor in terms of providing a rationale. Instances of explanatory behaviors are found in the following: explanations of value positions, i.e., Explain why you like Little Red Ridinghood best; conclusions, i.e., Explain why the wolf wanted to eat Little Red Ridinghood; main ideas, What is the main idea of this story?

EVALUATION. Evaluation questions deal with matters of value rather than matters of fact or inference and are thus characterized by their judgmental quality (worth, acceptability, probability, etc.) The following components of this category are adapted from a classification scheme by Aeschner and Gallagher (1965).

Questions calling for a rating (good, bad, true, etc.) on some item (idea, person, etc.) in terms of some scale of values provided by the teacher, i.e., Do you think this was a good or bad story?

Questions calling for a value judgment on a dimension set up by the teacher. Generally, these are "yes" or "no" responses following questions such as "Would you have liked Tom for a brother?"

Questions that develop from conjectural questions when the question is qualified by probability statements such as "most likely." "Do you think that it is most likely or least likely?"

Questions that present the pupil with a choice of two or more alternatives and require a choice, i.e., "Who did the better job in your opinion, Mary or Susan?"

APPENDIX D

THE BARRETT TAXONOMY--COGNITIVE AND AFFECTIVE DIMENSIONS OF READING COMPREHENSION[1]

1.0 Literal Comprehension. Literal comprehension focuses on ideas and information which are explicitly stated in the selection. Purposes for reading and teacher's questions designed to elicit responses at this level may range from simple to complex. A simple task in literal comprehension may be the recognition or recall of a single fact or incident. A more complex task might be the recognition or recall of a series of facts or the sequencing of incidents in a reading selection. Purposes and questions at this level may have the following characteristics.

1.1 Recognition requires the student to locate or identify ideas or information explicitly stated in the reading selection itself or in exercises which use the explicit ideas and information presented in the reading selection. Recognition tasks are:

1.11 Recognition of Details. The student is required to locate or identify facts such as the names of characters, the time of the story, or the place of the story.

1.12 Recognition of Main Ideas. The student is asked to locate or identify an explicit statement in or from a selection which is a main idea of a paragraph or a larger portion of the selection.

1.13 Recognition of a Sequence. The student is required to locate or identify the order of incidents or actions explicitly stated in the selection.

1.14 Recognition of Comparison. The student is requested to locate or identify likenesses and differences in characters, times, and

---

[1]T. C. Barrett. In H. M. Robinson (Ed.), Innovation and Change in Reading Instruction, (Chicago: National Society for the Study of Education, 1968), pp. 19-23).

places that are explicitly stated in the selection.

1.15 Recognition of Cause and Effect Relationships. The student in this instance may be required to locate or identify the explicitly stated reasons for certain happenings or actions in the selection.

1.16 Recognition of Character Traits. The student is required to identify or locate explicit statements about a character which helps to point up the type of person he is.

1.2 Recall requires the student to produce from memory ideas and information explicitly stated in the reading selection. Recall tasks are:

1.21 Recall of Details. The student is asked to produce from memory facts such as the names of characters, the time of the story, or the place of the story.

1.22 Recall of Main Ideas. The student is required to state a main idea of a paragraph or a larger portion of the selection from memory, when the main idea is explicitly stated in the selection.

1.23 Recall of a Sequence. The student is asked to provide from memory the order of incidents or actions explicitly stated in the selection.

1.24 Recall of Comparisons. The student is required to call up from memory the likenesses and differences in characters, times, and places that are explicitly stated in the selection.

1.25 Recall of Cause and Effect Relationships. The student is requested to produce from memory explicitly stated reasons for certain happenings or actions in the selection.

1.26 Recall of Character Traits. The student is asked to call up from memory explicit statements about characters which illustrate the type of persons they are.

2.0 Reorganization. Reorganization requires the student to analyze, synthesize, and/or organize ideas or information explicitly stated in the selection. To produce the desired thought product, the reader may utilize the statements of the author verbatim or he may paraphrase or translate the author's statements. Reorganization tasks are:

2.1 Classifying. In this instance the student is required to place people, things, places, and/or events into categories.

2.2 Outlining. The student is requested to organize the selection into outline form using direct statements or paraphrased statements from the selection.

2.3 Summarizing. The student is asked to condense the selection using direct or paraphrased statements from the selection.

2.4 Synthesizing. In this instance, the student is requested to consolidate explicit ideas or information from more than one source.

3.0 Inferential Comprehension. Inferential comprehension is demonstrated by the student when he uses the ideas and information explicitly stated in the selection, his intuition, and his personal experience as a basis for conjectures and hypotheses. Inferences drawn by the student may be either convergent or divergent in nature and the student may or may not be asked to verbalize the rationale underlying his inferences. In general, then, inferential comprehension is stimulated by purposes for reading and teachers' questions which demand thinking and imagination that go beyond the printed page.

3.1 Inferring Supporting Details. In this instance, the student is asked to conjecture about additional facts the author might have included in the selection which would have made it more informative, interesting, or appealing.

3.2 Inferring Main Ideas. The student is required to provide the main idea, general significance, theme, or moral which is not explicitly stated in the selection.

3.3 Inferring Sequence. The student, in this case, may be requested to conjecture as to what action

or incident might have taken place between two explicitly stated actions or incidents, or he may be asked to hypothesize about what would happen next if the selection had not ended as it did but had been extended.

3.4 Inferring Comparisons. The student is required to infer likenesses and differences in characters, times, or places. Such inferential comparisons revolve around ideas such as: "here and there," "then and now," "he and he," "he and she," and "she and she."

3.5 Inferring Cause and Effect Relationships. The student is required to hypothesize about the motivations of characters and their interactions with time and place. He may also be required to conjecture as to what caused the author to include certain ideas, words, characterizations, and actions in his writing.

3.6 Inferring Character Traits. In this case, the student is asked to hypothesize about the nature of characters on the basis of explicit clues presented in the selection.

3.7 Predicting Outcomes. The student is requested to read an initial portion of the selections and on the basis of this reading he is required to conjecture about the outcome of the selection.

3.8 Interpreting Figurative Language. The student, in this instance, is asked to infer literal meanings from the author's figurative use of language.

4.0 Evaluation. Purposes for reading and teacher's questions, in this instance, require responses by the student which indicate that he has made an evaluative judgment by comparing ideas presented in the selection with external criteria provided by the teacher, other authorities, or other written sources, or with internal criteria provided by the reader's experiences, knowledge, or values. In essence evaluation deals with judgment and focuses on qualities of accuracy, acceptability, desirability, worth, or probability of occurrence. Evaluative thinking may be demonstrated by asking the student to make the following judgments.

4.1 Judgments of Reality or Fantasy. Could this really happen? Such a question calls for a

judgment by the reader based on his experience.

4.2 Judgments of Fact or Opinion. Does the author provide adequate support for his conclusions. Is the author attempting to sway your thinking? Questions of this type require the student to analyze and evaluate the writing on the basis of the knowledge he has on the subject as well as to analyze and evaluate the intent of the author.

4.3 Judgments of Adequacy and Validity. Is the information presented here in keeping with what you have read on the subject in other sources? Questions of this nature call for the reader to compare written sources of information, with an eye toward agreement and disagreement or completeness and incompleteness.

4.4 Judgments of Appropriateness. What part of the story best describes the main character? Such a question requires the reader to make a judgment about the relative adequacy of different parts of the selection to answer the question.

4.5 Judgments of Worth, Desirability and Acceptability. Was the character right or wrong in what he did? Was his behavior good or bad? Questions of this nature call for judgments based on the reader's moral code or his value system.

5.0 Appreciation. Appreciation involves all the previously cited cognitive dimensions of reading, for it deals with the psychological and aesthetic impact of the selection on the reader. Appreciation calls for the student to be emotionally and aesthetically sensitive to the work and to have a reaction to the worth of its psychological and artistic elements. Appreciation includes both the knowledge of and the emotional response to literary techniques, forms, styles, and structures.

5.1 Emotional Response to the Content. The student is required to verbalize his feelings about the selection in terms of interest, excitement, boredom, fear, hate, amusement, etc. It is concerned with the emotional impact of the total work on the reader.

5.2 Identification with Characters or Incidents. Teachers' questions of this nature will elicit responses from the reader which demonstrate his

sensitivity to, sympathy for, and empathy with characters and happenings portrayed by the author.

5.3 Reactions to the Author's Use of Language. In this instance the student is required to respond to the author's craftsmanship in terms of the semantic dimensions of the selection, namely, connotations and denotations of words.

5.4 Imagery. In this instance, the reader is required to verbalize his feelings with regard to the author's artistic ability to paint word pictures which cause the reader to visualize, smell, taste, hear, or feel.

APPENDIX E

TAXONOMY OF READING COMPREHENSION[1]

1.0 Literal Recognition or Recall. Literal comprehension requires the recognition or recall of ideas, information, and happenings that are explicitly stated in the materials read. Recognition Tasks, which frequently take the form of purposes for reading, require the student to locate or identify explicit statements in the reading selection itself or in exercises that use the explicit content of the reading selection. Recall tasks demand the student to produce from memory explicit statements from a selection; such tasks are often in the form of questions teachers pose to students after a reading is completed. Two additional comments seem warranted with regard to literal comprehension tasks. First, although literal comprehension tasks can be overused, their importance cannot be denied, since a student's ability to deal with such tasks is fundamental to his ability to deal with other types of comprehension tasks. Second, all literal comprehension tasks are not necessarily of equal difficulty. For example, the recognition or recall of a single fact or incident may be somewhat easier than the recognition or recall of a number of facts or incidents, while a more difficult task than either of these two may be the recognition or recall of a number of events or incidents and the sequence of their occurrence. Also related to this concern is the hypothesis that a recall task is usually more difficult than a recognition task, when the two tasks deal with the same content and are of the same nature. Some examples[2] of literal comprehension tasks are:

---

[1] T. C. Barrett. In R. J. Smith & T. C. Barrett, Teaching reading in the middle grades (Reading, Mass.: Addison-Wesley, 1974), pp. 58-57).

[2] Although the examples in each of the categories are logically ordered from easy to difficult, it is recognized that such a finite hierarchy has not been validated. Therefore, the user of the Taxonomy should view the examples as some of the tasks that might be used to help students produce comprehension products that relate to the type of comprehension described in each of the four major categories of the Taxonomy.

1.1 Recognition or Recall of Details. The student is required to locate or identify or to call up from memory such facts as the names of characters, the time a story took place, the setting of a story, or an incident described in a story, when such facts are explicitly stated in the selection.

1.2 Recognition or Recall of Main Ideas. The student is asked to locate or identify or to produce from memory an explicit statement in or from a selection which is the main idea of a paragraph or a larger portion of the selection.

1.3 Recognition or Recall of Sequence. The student is required to locate or identify or to call up from memory the order of incidents or actions explicitly stated in the selection.

1.4 Recognition or Recall of Comparisons. The student is requested to locate or identify or to produce from memory likenesses and differences among characters, times in history, or places that are explicitly compared by an author.

1.5 Recognition or Recall of Cause and Effect Relationships. The student in this instance may be required to locate or identify or to produce from memory reasons for certain incidents, events, or characters' actions explicitly stated in the selection.

1.6 Recognition or Recall of Character Traits. The student is requested to identify or locate or to call up from memory statements about a character which help to point up the type of person he was when such statements were made by the author of the selection.

2.0 Inference. Inferential comprehension is demonstrated by the student when he uses a synthesis of the literal content of a selection, his personal knowledge, his intuition and his imagination as a basis for conjectures or hypotheses. Conjectures or hypotheses derived in this manner may be along convergent or divergent lines, depending on the nature of the task and the reading materials involved. For example, inferential tasks related to narrative selections may permit more divergent or creative conjectures because of the open-ended possibilities provided by such writing. On the other hand, expository selections, because of their content, may call for convergent hypotheses more

often than not. In either instance, students may or may not be called upon to indicate the rationale underlying their hypotheses or conjectures, although such a requirement would seem to be more appropriate for convergent rather than divergent hypotheses. Generally, then, inferential comprehension is elicited by purposes for reading, and by teachers' questions which demand thinking and imagination which are stimulated by, but go beyond, the printed page. Examples of inferential tasks related to reading are:

2.1 Inferring Supporting Details. In this instance, the student is asked to conjecture about additional facts the author might have included in the selection which would have made it more informative, interesting or appealing.

2.2 Inferring the Main Idea. The student is required to provide the main idea, general significance, theme, or moral which is not explicitly stated in the selection.

2.3 Inferring Sequence. The student, in this case, may be requested to conjecture as to what action or incident might have taken place between two explicitly stated actions or incidents; he may be asked to hypothesize about what would happen next; or he may be asked to hypothesize about the beginning of a story if the author had not started where he did.

2.4 Inferring Comparisons. The student is required to infer likenesses and differences in characters, times, or places. Such inferential comparisons revolve around ideas such as: "here and there," "then and now," "he and he," "he and she," and "she and she."

2.5 Inferring Cause and Effect Relationships. The student is required to hypothesize about the motives of characters and their interactions with others and with time and place. He may also be required to conjecture as to what caused the author to include certain ideas, words, characterizations, and actions in this writing.

2.6 Inferring Character Traits. In this case, the student may be asked to hypothesize about the nature of characters on the basis of explicit clues presented in the selection.

2.7 Predicting Outcomes. The student is requested to read an initial portion of selection, and on the basis of this reading to conjecture about the outcome of the selection.

2.8 Inferring about Figurative Language. The student, in this instance, is asked to infer literal meanings from the author's figurative use of language.

3.0 Evaluation. Evaluation is demonstrated by a student when he makes judgments about the content of a reading selection by comparing it with external criteria, e.g., information provided by the teacher on the subject, authorities on the subject, or by accredited written sources on the subject; or with internal criteria, e.g., the reader's experiences, knowledge, or values related to the subject under consideration. In essence, evaluation requires students to make judgments about the content of their reading, judgments that have to do with its accuracy, acceptability, worth, desirability, completeness, suitability, timeliness, quality, truthfulness, or probability of occurrence. Examples of evaluation tasks related to reading are:

3.1 Judgments of Reality or Fantasy. The student is requested to determine whether incidents, events, or characters in a selection could have existed or occurred in real life on the basis of his experience.

3.2 Judgments of Fact or Opinion. In this case the student is asked to decide whether the author is presenting information which can be supported with objective data or whether the author is attempting to sway the reader's thinking through the use of subjective content that has overtones of propaganda.

3.3 Judgments of Adequacy or Validity. Tasks of this type call for the reader to judge whether the author's treatment of a subject is accurate and complete when compared to other sources on the subject. In this instance, then, the reader is called upon to compare written sources of information with an eye toward their agreements or disagreements, their completeness or incompleteness, and their thoroughness or superficiality in dealing with a subject.

3.4 Judgments of Appropriateness. Evaluation tasks of this type require the student to determine whether certain selections or parts of selections are relevant and can contribute to resolving an issue or a problem. For example, a student may be requested to judge the part of a selection which most appropriately describes a character. Or he may be called upon to determine which references will make significant contributions to a report he is preparing.

3.5 Judgments of Worth, Desirability, or Acceptability. In this instance, the student may be requested to pass judgments on the suitability of a character's action in a particular incident or episode. Was the character right or wrong, good or bad, or somewhere in between? Tasks of this nature call for opinions based on the values the reader has acquired through his personal experiences.

4.0 Appreciation. Appreciation has to do with students' awareness of the literary techniques, forms, styles, and structures employed by authors to stimulate emotional responses in their readers. Obviously, tasks which fall into this category will require varying degrees of inference and evaluation, but their primary focus must be on heightening students' sensitivity to the ways authors achieve an emotional as well as an intellectual impact on their readers. More specifically, appreciation involves cognizance of and visceral response to: (a) the artistry involved in developing stimulating plots, themes, settings, incidents, and characters, and (b) the artistry involved in selecting and using stimulating language, in general. Examples of tasks that involve appreciation are:

4.1 Emotional Response to Plot or Theme. Tasks of this type are based on the assumption that the plot or the theme of a given selection has stimulated and sustained a feeling of fascination, excitement, curiosity, boredom, sentimentality, tenderness, love, fear, hate, happiness, cheerfulness or sadness. Provided this assumption is met, the students may be requested to determine what the author did in the process of developing the plot or theme that elicited a given emotional response.

4.2 Identification with Characters and Incidents.

Some appreciation tasks should require students to become aware of the literary techniques and devices which prompt them to sympathize or empathize with a particular character, or to reject him, for that matter. Other tasks should require students to consider the placement, nature, and structure of events or incidents which cause them to project themselves into the action.

4.3 Reactions to the Author's Use of Language. In this instance, the student is required to recognize and respond to the author's craftsmanship as reflected in his selection of and use of words. Such tasks may deal with the connotations and denotations of selected words and the influence they have on a reader's feelings. In addition, students should at times note figures of speech, e.g., similes and metaphors, and the effect their use has on the reader.

4.4 Imagery. Tasks of this nature require the reader to recognize and react to the author's artistic ability to "paint word pictures." In other words, students should become sensitive to the techniques an author uses in order to enable them to see, smell, taste, hear, or feel things through reading.

APPENDIX F

SANTOS' CLASSIFICATION SCHEME FOR READING QUESTIONS[1]

1. Recognition or Recall Questions

A reading question falls under this category if it can be answered verbatim with specific information found in the selection read, and if it requires the reader to repeat or locate the information, as in:

a. responding to direct factual questions in which the fact is given in the selection read;

b. enumerating or making a list of information presented in the selection read;

c. quoting the author or character in the selection;

d. recalling who said a given quotation taken verbatim from the selection read;

e. reciting a stanza of a poem, an essay, rules, from memory;

f. reading aloud the part that answers a factual question, or the part called for.

2. Translation Questions

A reading question falls under this category if it requires the reader to restate a specific information in his own words or in another form of communication, while keeping the idea basically unchanged, as in:

a. explaining in his own words the meaning of technical terms, vocabulary words, and stated meanings;

b. restating a problem, generalization, principle, rule, procedure, in his own words;

c. paraphrasing parts of a selection, or summarizing

---

[1] I. Santos. Cited in H. W. Forgan & C. T. Mangrum II, Teaching Content Area Reading Skills: A Modular Preservice and Inservice Program (Columbus, Ohio: Merrill, 1976), pp. 176-183.

a selection;

d. retelling a selection according to a given set of topics; describing or retelling parts of events of a selection to prove a point; retelling additional stories related to the topic in his own words;

e. following printed directions;

f. changing poetry to prose or vice versa; changing a statement to a question;

g. illustrating or drawing information recorded verbally; describing a picture, poster, sketch, drawing, painting, or object in his own words;

h. presenting information by making a graph, diagram, chart, table, or map.

3. Application Questions

A reading question falls under this category if it requires the reader to recall an appropriate and previously learned skill or abstraction (generalization, principle, theory, rule, technique, or procedure) and then use it to explain or solve a problematic situation which is presented in a new context, as in:

a. making use of previous knowledge to make comparisons, observations, and explanations, or to answer a question;

b. applying new information or principles learned from the selection read to explain a phenomena or a problematic situation;

c. using a theory or principle to predict the probable effects of certain factors;

d. applying rules or organizing ideas, as in:

(1) classifying or grouping

(2) indexing

(3) outlining

(4) rearranging a given set of ideas alphabetically, chronologically, or in proper sequence;

e. applying arithmetical procedures and skills to solve and/or explain arithmetical problems;

f. applying skill in locating information, as in:

(1) a dictionary

(2) an encyclopedia

(3) an atlas

(4) other general references;

g. identifying and/or giving examples to fit a previously-stated definition or abstraction;

h. identifying elements of style used in the selection read, such as rhyme, free verse, satire;

i. identify propaganda devices used in persuasive material, facts vs. opinions, author's biases or prejudices, and emotive language;

j. identifying literary devices of humor, wit, satire, mood, tone, and figures of speech.

4. Analysis Questions

A reading question falls under this category if it requires the reader to identify the component parts of a given communication, and/or to determine relationships among these component parts or among parts of ideas as in:

a. indicating which are the assumptions, conclusions, generalizations, or supporting facts explicit in a given communication;

b. making inferences, conclusions, generalizations, or implications; anticipating or predicting outcomes from evidence presented;

c. identifying in a given communication what information is given, what is not given, and/or what is to be solved in a story problem;

d. identifying clues from which inferences may be drawn, as in:

(1) connotative expressions

(2) character's actions, traits, remarks

(3) the setting

(4) figurative language

(5) semantic variations;

e. identifying relationships, as in:

(1) cause-effect relationships

(2) sequential relationships

(3) logical relationships (analogies, antecedents)

(4) past-present relationship

(5) main ideas and details

(6) central theme;

f. indicating whether ideas are similar, different, contradictory, supportive, related, relevant, irrelevant, essential, extraneous, in proper sequence, true or false, right or wrong, good or bad;

g. identifying author's purpose and point of view;

h. identifying parts of a given communication which makes the communication false.

5. Synthesis Questions

A reading question falls under this category if it offers a variety of possibilities and allows the reader freedom of expression. At the same time it requires him to compose a unique communication, devise a proposed procedure, or derive new relationships, as in:

a. using a given word, phrase, or expression in his own words;

b. rearranging a given set of ideas into unique sentences or paragraphs; substituting words or phrases to make ideas more correct;

c. writing an original rhyme, poem, essay, or story;

d. reading aloud the part called for and giving the passage his own expression; dramatic reading; choral reading;

e. pantomiming; retelling or dramatizing a selection or parts of it using one's own words and expressions; presenting a pageant;

f. relating a personal experience;

g. discussing or giving a short talk or extemporaneous speech on a given topic; telling an original story;

h. submitting a plan for a project needing construction, an experiment, a program, a skit, or a dramatization;

i. suggesting solutions to a problem; listing possible problems that might sum up a situation;

j. suggesting different titles to a given communication; suggesting scenes or events pictured from a given communication;

k. formulating rules to observe in performing or judging a performance or presentation;

l. giving different endings to a given communication;

m. relating a general problem to a local situation;

n. sharing personal experiences; relating one's personal experience to the content of what is read or to the question asked;

o. pretending;

p. giving an opinion or making decisions without justifying one's views.

6. Evaluation Questions

A reading question falls under this category if it requires the reader to determine whether a given communication meets standards set up by the student herself or presented to her. It also requires her to make decisions and to support her views in her own words, as in:

a. judging agreement between ideas or sources of information, and justifying one's views;

b. forming an opinion or taking a stand on an issue, and justifying one's views;

c. indicating whether a conclusion, generalization,

main theme, or thesis logically follows the facts presented; selecting the best conclusion from a set of suggested conclusions, and justifying one's views;

d. indicating whether ideas are similar, different, contradictory, supportive, related, relevant, essential, or in proper sequence, and justifying one's views.

e. judging or criticizing a given communication, and justifying one's views for:

accuracy	completeness
authenticity	practicability
relevancy	reasonableness
adequacy	authoritativeness
value	logical consistency
imagination	etc.;

f. judging the relative merits of information from different sources, or sources of information, and justifying one's views;

g. judging characters in a story, and justifying one's views;

h. evaluating suggested endings to a story, summaries, a performance, a declamation, an oration, a speech, etc., and justifying one's views and/or suggesting ways of improvement;

i. evaluating pictures for accuracy, appropriateness, suitability, richness of imagination, and justifying one's views;

j. evaluating statements for ambiguities, overspecificity, omissions, distortions, inconsistencies, inaccuracies, overstatements, and justifying one's views;

k. judging and proving the truth or falsity of statements by pointing out supportive statements or evidence, and justifying one's views.

APPENDIX G

CLASSIFICATION OF TEST ITEMS ACCORDING TO BARRETT'S TAXONOMY OF READING COMPREHENSION

Worksheet A-1

Test: CAT (#1-42) Judge A

Recognition & Recall	Inference	Evaluation	Appreci-ation	Problem Items
1	3	35		
2	7	36		
4	11	37		
5	12	38		
6	13	39		
8	15	40		
9	16	41		
10	17	42		
14	19			
18	20			
26	21			
28	22			
29	23			
32	24			
	25			
	27			
	30			
	31			
	33			
	34			
Totals:				
14	20	8	0	0

Worksheet A-2

Test: CTBS (#1-45) Judge A

Recognition & Recall	Inference	Evaluation	Appreci-ation	Problem Items[a]
1	2	30[a]	3	
6	7		4	
13	8		5	19
14	9		18	30
23	10			38
26	11			
33	12			
35	15			
40	16			
45	17			
	19[a]			
	20			
	21			
	22			
	24			
	25			
	27			
	28			
	29			
	31			
	32			
	34			
	36			
	37			
	38[a]			
	39			
	41			
	42			
	43			
	44			
Totals:				
10	30	1	4	3

[a]All problem items were finally placed and counted in a taxonomic category.

Worksheet A-3

Test: ITBS (#25-98) Judge A

Recognition & Recall	Inference		Evaluation	Appreci-ation	Problem Items
25	27	65		59[a]	30
26	28	67		60	38
31	29	68		91	40
32	30[a]	69		93	56
33	34	70[a]		94	59
37	35	71		96	97
40[a]	36	72		97[a]	98
57	38[a]	73			
62	39	74			
64	41	75			
66	42	76			
	43	77			
	44	78			
	45	79			
	46	80			
	47	81			
	48	82			
	49	83			
	50	84			
	51	85			
	52	86			
	53	87			
	54	88			
	55	89			
	56[a]	90			
	58	92			
	61	98[a]			
	63				
Totals:					
11	55		0	8	8

[a]All problem items were finally placed and counted in a taxonomic category.

Worksheet A-4

Test: MAT (#1-45) Judge A

Recognition & Recall	Inference		Evaluation	Appreci-ation	Problem Items
13	1	24[a]		35	
	2	25			5
	3	26			
	4	27[a]			11
	5[a]	28			
	6	29			18
	7	30			24
	8	31[a]			27
	9	32			31
	10	33			41
	11[a]	34			
	12	36			
	14	37			
	15	38			
	16	39			
	17	40			
	18[a]	41[a]			
	19	42			
	20	43			
	21	44			
	22	45			
	23				
Totals:					
1	43		0	1	7

[a]All problem items were finally placed and counted in a taxonomic category.

Worksheet A-5

Test: SRA (#1-60) Judge A

Recognition & Recall	Inference		Evaluation	Appreci-ation	Problem Items
9	1	32[a]	14	7	
10	2	33	39		6
44	3	34			19
49	4	35			32
	5	36			
	6[a]	37			
	8	38			
	11	40			
	12	41			
	13	42			
	15	43			
	16	45			
	17	46			
	18	47			
	19[a]	48			
	20	50			
	21	51			
	22	52			
	23	53			
	24	54			
	25	55			
	26	56			
	27	57			
	28	58			
	29	59			
	30	60			
	31				
Totals:					
4	53		2	1	3

[a]All problem items were finally placed and counted in a taxonomic category.

Worksheet A-6

Test: STEP (#1-30) Judge A

Recognition & Recall	Inference	Evaluation	Appreciation	Problem Items
12	1		8	
13	2			25
	3			
	4			
	5[a]			
	6			
	7			
	9			
	10			
	11			
	14			
	15			
	16			
	17			
	18			
	19			
	20			
	21			
	22			
	23			
	24			
	25[a]			
	26			
	27			
	28			
	29			
	30			
Totals:				
2	27	0	1	2

[a]All problem items were finally placed and counted in a taxonomic category.

Worksheet A-7

Test: SAT (#1-71) Judge A

Recognition & Recall	Inference		Evaluation	Appreci-ation	Problem Items
1	2	37			4
3	4[a]	38			12
15	5	39			15
20	6	40			22
26	7	41			31
46	8	42			36
50	9	43			44
51	10	44[a]			54
62	11	45			64
	12[a]	47			71
	13	48			
	14	49			
	15[a]	52			
	16	53			
	17	54			
	19	55			
	21	56			
	22[a]	57			
	23	58			
	24	59			
	25	60			
	27	61			
	28	63			
	29	64[a]			
	30	65			
	31[a]	66			
	32	67			
	33	68			
	34	69			
	35	70			
	36[a]	71[a]			
Totals:					
9	62		0	0	10

[a]All problem items were finally placed and counted in a taxonomic category.

APPENDIX H

PROBLEM ITEMS: REASONS FOR THE DIFFICULTY OF PLACEMENT--JUDGE A

Test	Item	Asks for "best" answer	Other Difficulties
CTBS	19	X	Requires judgment among Inferences about character
	30	X	Must use criteria of personal values
	38	X	
ITBS	30	X	Vocabulary item; requires judgment based on experience, but student must rule out figurative use of word
	38	X	No category available for translation to visual form, but it is essentially an Inference task
	40	X	Could also be Inference
	56		Vocabulary item
	59	X	Could be Inference, but goes beyond interpreting figurative language to expressing identification with subject of poem
	70	X	
	97	X	Also requires some reader sensitivity to poet's situation

PROBLEM ITEMS: REASONS FOR THE DIFFICULTY OF PLACEMENT--JUDGE A

Test	Item	Asks for "best" answer	Other Difficulties
ITBS	98	X	Could test reader's sensitivity to theme, but is essentially concerned with inferring possible outcome of situation
MAT	5	X	
	11	X	
	18	X	
	24	X	
	27		Vocabulary item
	31	X	
	41	X	Actually focuses on inferring character traits
SRA	6		Asks for a "good" answer
	19	X	
	32	X	Actually focuses on inferring character traits

PROBLEM ITEMS: REASONS FOR THE DIFFICULTY OF PLACEMENT--JUDGE A

Test	Item	Asks for "best" answer	Other Difficulties
SAT	4	X	
	12	X	
	16	X	
	22	X	
	31	X	
	36	X	
	44		Pronoun reference; requires Inference
	54		Factual, but not a verbatim statement; requires Inference
	64	X	
	71	X	
STEP	5	X	
	25	X	

APPENDIX I

SUMMARY OF PLACEMENT OF SAMPLE OF ITEMS
SHOWING ITEM NUMBERS--JUDGE B

Test	Recognition	Inference	Evaluation	Appreciation
CAT	9			
	12			
	16			
	24			
	38			
CTBS	14	9		
		31		
		40		
		45		
ITBS	47	25	58	
		87	72	
MAT	1	23		
	12	33		
		36		
SRA	25	2	29	
		30		
		48		
STEP	7	28	30	17
		29		
SAT	3	36		
	8	64		
	39			
Total	f=14 %=40.0	f=16 %=45.7	f=4 %=11.4	f=1 %=2.8

APPENDIX J

PROBLEM ITEMS: REASONS FOR DIFFICULTY OF PLACEMENT--JUDGE B

Test	Item	Reason for Difficulty
CAT	16	Called literal, although it requires slight translation
CTBS	9	A vocabulary item. Answer could be inferred from context. Vocabulary category is not provided under Inference.
CTBS	40	Student asked to identify antecedent; structure is ambiguous. Similar to Category 2.8
ITBS	72	Item requires application of knowledge not given in the story
MAT	12	A translation item, possibly inferential
SRA	29	No sub-category fits. Item requires evaluation based on a definition that must be made by analogy to a selection of the passage, yet no evaluation sub-category identifies this type of question.
SRA	40	Very close to literal recognition. Item requires translation of ideas or some synthesis, therefore, included under Inference. Information in selection is not explicit
STEP	29	Classified Inferential. Although knowledge of definition of key term is required, other phrases in the paragraph act as synonyms